LET'S VACATION!

Caleb Carlson, Aislinn Jeske, and Cassidy Skorczewski

May 3, 2019

Contents

1 Introduction

It is a generally accepted fact that leisure travelling and vacation positively affect mental health, well-being, and hope for an individual's future [1]. Over the span of a career, adults may experience job burnout and fatigue due to repetitive tasks being performed or too many hours worked under a high stress load or lack of social interaction. This burnout can be a cause that lead individuals into anti-social tendencies and even depression. However, something as simple as a short trip to can relieve these symptoms [2]. Personally, I have noticed that after weeks in the computer lab, even just a trip to the mountains drastically heightens my sense of the "big picture" and hope for my future.

Since vacationing is crucial for career stamina, it logically follows that trip planning should add minimal or no stress to the life of the potential traveller. Unfortunately, a study by Jeroen Nawijn shows that the period of time leading up to a vacation is "more of a burden than a blessing" [3]. Some of this stress can be attributed to the "unknown" aspects of the trip, along with the work required to research and plan it. Services such as Kayak.com aim to reduce the workload of trip planning by finding and suggesting the cheapest elements of a trip for a given destination. For example, if one wished to travel to Seattle, Washington, Kayak.com could filter out suboptimal care rentals and flight rates, finding just the ones which are the "best bang for the buck". This service is now widely used in place of the older-style travel agents, making trip planning much more efficient and affordable.

This paper attempts to take this research even one step further, and explore the possibility of filtering destinations for a user based on a budget. Our research question is as follows: Given a budget, can we determine a set of global destination possibilities along with a reasonable length-of-stay for the user? Below, we explore the methodology used in our project to filter global data and determine a cost-per-day rate at any given destination on a high, medium, or low budget. We benchmark the daily cost of popular destinations against each other to compare results against common sense (i.e. Switzerland should be more expensive than Brazil). Next, we draw a conclusion from the results, showing that it is indeed possible to accurately suggest destinations for a user under a budget. Finally, we speculate on future problem spaces that could evolve to require changes in our logic and process.

While the project itself uses a rudimentary approach to approximating daily cost, the underlying ideas and processes used contain the key to providing a user with an easy trip-planning experience, greatly reducing the amount of burden they may experience pre-vacation. With complex data querying, filtering, and estimation logic, we believe our approach to destination suggestion will almost always satisfy the needs of the user by decreasing the list of possible trips by orders of magnitude. From this significantly smaller list of destinations that fit a given budget, the user may easily choose the trip that best fits their definition of leisure.

2 Problem Characterization

In 2017, there were more than 1.326 million international tourist arrivals around the world [4]. Each of these travellers had to make decisions on what hotel they would stay in, what they would eat, what activities they would do and places they would visit, but most importantly, where they would travel. This decision can be difficult and limiting, especially when that person has a tight budget. Unfortunately this problem is not an easy one for the average traveller to solve because there is a profusion of data for some aspects of travelling, such as lodging costs, and a sparse amount of data for others, such as food costs.

The first problem travellers face is deciding where their destination will be. Usually people have an idea in their mind of where they want to go in the world, but this dream may fall short for people with a tight budget because some places in the world are more expensive than others. It is infeasible

for someone to calculate exactly how far their budget would take them in dozens of places around the world, but it is an important step when planning a trip because it is not worth the money to fly across the world to stay 2 days at a destination. Though it is an important step for any traveller, another challenge arises when a traveller attempts to navigate a website with copious amounts of data.

In the age where information is more accessible than ever before, the power of all this information can be lost on the average user if they do not know how to navigate it effectively. With the large number of hotel and hostel sites such as Airbnb, it is hard for a traveler to know if they are maximizing their stay with the housing option they chose. There are over 20,000 Airbnb housing options in Amsterdam alone [5]. This is part of what makes this problem so challenging; that amount of data can be overwhelming for the average traveller. Without the background knowledge on how to reduce this data to get the best destination for their budget, the traveller is limited to places they can research themselves.

After the flights and hotels are booked, travel costs do not end there. Once they have arrived at their destination, food and transportation still need to be paid for. This poses a difficult problem for budget travellers because the exact cost of food and transportation is unknown before the trip begins. A traveller may not even have an idea of how much they are going to spend on these two factors and may not be able to budget correctly. It is impossible to get an exact value of how much it is going to cost for the whole trip, but there is data out there that estimates these costs for specific cities around the world. Although this data exists on sites like Numbeo, the average traveller may not know to look there for this information. Additionally, this data is sparse for this kind of information, even Numbeo gets their data from locals who input average costs for particular food items, transportation types, and leisure activities. It is impossible to know how much it is going to cost without planning out activities, bars, and restaurants exactly, down to the meal and tip. The best a traveller can do is estimate.

3 Dominant Approaches to the Problem

A possible approach to this problem can be best explained through the existing website kayak.com. Kayak is a metasearch engine that collects travel information from multiple websites and displays it in one place. The main advantage of this method is that it provides information from multiple websites for hotels, flights, and rental cars and is easily expandable to include other budget friendly options. It does this by compiling prices from other sites and displaying it on their own website, referring users to a link when they wish to book a hotel or flight [6]. This allows for high efficiency and accuracy because the prices come directly from the third party website. It also allows for smaller overhead for kayak.com because no transactions occur directly on their site. Another advantage is that this basic idea can be expanded for more budget friendly options such as Airbnb, HostelWorld, or other low budget housing options. The final advantage of this method is that it provides up to date information, while a dataset may contain some stale data. The main disadvantage of this method involves the scope of information it provides. Although there is a multitude of information for some aspects of travelling, it does not include any information for the costs once a traveller arrives at their destination. Because of this, after using kayak.com, for example, users are left with a good idea of how much they are spending on flights, hotels, and possibly rental cars, but no idea of how much the rest of the trip is going to cost them. As stated in Section 2, this information is difficult to come across, but important for a possible traveller to know. This method of solving the problem provides a single source for the major components for planning a trip for the user, with little oversight by the company.

4 Methodology

As previously stated, the big-picture goal of the project is to provide the user with a list of possible destinations, along with the durations they could potentially stay with their specific budget. The most straightforward approach to computing these durations is to first find the cost-per-day at a given destination, under three different budget levels. Because the individual's lifestyle is unknown to the application, these three budget levels were chosen to roughly estimate a "budget" lifestyle (possibly that of a college student), an average "middle-class" lifestyle matching that of a young adult or entrepreneur, and lastly a "wealthy" budget. Thus, the user can use the budget level which most closely matches their own. Complex approaches can be taken to accurately estimate a daily cost on vacation, but due to time constraints, the project had to be limited to just three factors: nightly lodging, food, and transportation expenses. It is thought that, excluding airfare, these are the most significant and regular expenses to be incurred during a vacation.

4.1 City Pricing

In order to calculate the average cost of food per day that a user would consume, we needed to find a global dataset that was frequently updated and also allowed us to pinpoint the costs of restaurants globally on a city-by-city basis. The first method that came to mind to do this was Google. We knew that Google listed price indicators (denoted $ to $$$$) along with almost every restaurant listed on its Map service. Google states that each "$" represents a $10 increment in price. Not only did this price indication fit nicely with our 3-tier budget model, but Google extended this price indication feature globally to restaurants worldwide. Unfortunately, we discovered that Google most likely sourced its price indication levels from Yelp reviews through a proprietary algorithm, and did not offer the data for public use.

Luckily, we stumbled across Numbeo, an even better solution to our problem. Numbeo is a service that provides cost-of-living data worldwide for a massive selection of services and products, is frequently updated, and provides an API for querying their database. Much to our surprise, average restaurant, market, and transportation costs had already been computed for us for different types of restaurants, food items, and modes of transport on a city-by-city basis. It only allowed for querying one city at a time, so to remedy this, we created a program that would read city names out of a file, and query the database for each city read in. The results of a query get returned as a JSON object, which could be easily parsed into a Java object and added to a list of "City" objects with their respective pricing data. Finally, we had the program write every city and its pricing data to a file in CSV format, which is one of the most standard input formats of Spark.

With all the city pricing data in a CSV file on the Hadoop Distributed File System (HDFS), we discussed what items would constitute each budget level for each city, and dropped the irrelevant columns from the dataframes in Spark. The items we chose for each budget level are displayed below, along with the rationale behind each decision.

4.1.1 Expenses per day

Low Budget, Food: 2x Low-tier (McDonalds or equivalent) restaurant, 1x homemade sandwich with ingredients bought from market: 1/3 Lettuce Head, 1/3 Tomato, 1/7 Chicken Breast, 1/3 Loaf of Bread.

Low Budget, Transport: 2x One-way Bus Tickets

Mid Budget, Food: 1x Low-tier (McDonalds or equivalent) restaurant, 2x Mid-tier restaurant, 1x

Domestic Beer.

Mid Budget, Transport: 10km of Taxi or equivalent Uber transport.

High Budget, Food: 2x Mid-tier restaurant, 1x High-tier restaurant (3-course meal), 1x Domestic Beer.

High Budget, Transport: 20km of Taxi or equivalent Uber transport.

4.1.2 Rationale

We assumed that if a traveller wanted to stay on a budget, they would most likely eat at cheaper fast-food restaurants and make their own meals with ingredients from local markets. Furthermore, a minimum of two one-way bus tickets would usually be required to accomplish a daily activity not within walking distance of their destination. On a middle-class budget, we speculated that the user would not be making their own food, and might be taking Ubers or Taxis to sight-see. Lastly, we speculated that wealthier travellers would be doing multiple activities in a day, and would be treating themselves to classier dining more often than not.

After applying these mappings to the Spark RDD columns, we observed that the results for each city's daily expenses seemed absolutely reasonable for the individual budget levels. For example, it was noted that the low, mid, and high daily expenses for Austin, Texas were $\sim$ \$22, \$35, and \$60 respectively.

4.2 Lodging Pricing

In order to calculate lodging expenses, we turned to Airbnb. Airbnb is a service in which someone lists their home, or a room in their home, online for travellers to rent when visiting a new city. Airbnb services over 191 countries and 81 thousand cities, for a total of 6 million listings worldwide [7]. Airbnb released housing data for 91 metropolis areas that could be used for analytical purposes. Each city was in its own CSV file that was loaded onto the HDFS to be analyzed using Spark. The CSV files had over 75 features for every Airbnb listing. We only analyzed the features that detailed the location, the room-type, price per night, and the fees associated with that listing. For each of the budget levels in every city, we found the average price-per-night costs and the average cleaning/damage fees. These are the values we used to estimate the lodging expenses in each city. The items we chose for each budget level are displayed below, along with the rationale behind each decision.

4.2.1 Expenses per day

Low Budget, Lodging: A Shared Room in a Metropolis Area

Mid Budget, Lodging: A Private Room in a Metropolis Area

High Budget, Lodging: An Entire Home or Apartment in a Metropolis Area

4.2.2 Rationale

We assumed that if a traveller wanted to stay on a budget, they would most likely want to rent as simple, and as small, as possible. The room-type options for Airbnb listings are: shared room,

private room, and an entire home/apt. Based on these options, it was the logical choice to make the lowest budget a shared room, the middle budget to be a private room, and a high budget to be an entire home/apt. After applying these mappings to the Spark RDD columns, we observed that the results for each city's lodging expenses seemed absolutely reasonable for the individual budget levels. For example, it was noted that the low, mid, and high lodging expenses for Denver, Colorado were $\sim$ \$41, \$68, and \$158 respectively.

4.3 Data Combination

Given the averages for both food and transportation and housing for all three budget types, it was a simple task to calculate the duration of stay for each city. The two types of expenses were read in from HDFS as different RDDs in Spark. This was necessary because certain filtering needed to be performed on the Airbnb data. Listings are posted by the hosts themselves and Airbnb does not perform quality checks, so some of the data was misspelled, or listed a city under the wrong country. The filtering effectively removed all cities that were incorrectly posted. The traveller's budget was entered into the program as a command line argument, so it could be run easily with multiple possible budgets. Once this was performed and all data was valid, the two RDDs were joined and grouped by key. In this case, the logical choice for the key was the city name and the budget type (lo, mid, hi). This grouped together food, transport, housing, and housing fees for a single city and budget type. Once this was completed, a simple calculation needed to be performed to determine how long the traveller could stay in a city. The equation used is as follows:

$$Duration = \frac{Budget - HousingFee}{HousingCost + Food + Transport} \tag{1}$$

5 Experimental Benchmarks

The metric we used to assess the quality of our solution was accuracy. It is common knowledge that some parts of the world are more expensive than others to live and vacation. One useful measurement that allows you to compare expenses in cities around the world is the cost of living index associated with each city [8]. Using this index, cities can be ranked from most expensive place to live, to the least expensive place to live. We can assess the accuracy of our solution by comparing specific cities to see if the number of days you can spend in those cities reflects their index ranking and the traveller's budget. We used this basis to test whether the travel recommendations are correct.

5.1 Varying Budget Levels

One of the ways we checked the accuracy of our solution was to compare the amount of days you could spend in each city at the different budget levels. We would expect that if you spent less money per day in a city, then you would be able to stay in that place longer than if you spent a higher amount of money per day. Using `python`, we randomly selected 20 of our 89 cities to examine their different budget levels as seen in Figure 1.

We decided to randomly select 20 cities instead of showing all the cities so our graph would not be as cluttered. Looking at all the cities in Figure 1, we can see that with the low budget, you can spend more days in every city than the mid budget and the high budget. Based on the Cost of Living Index mentioned previously, the Economist [9] states that of the cities we chose to analyze, Hong Kong, Copenhagen, Zurich, Paris, New York, and Vaud are the most expensive cities to live, and Prague, Malaga, Porto, and Lisbon are the least expensive cities to live. In Figure 1, Copenhagen

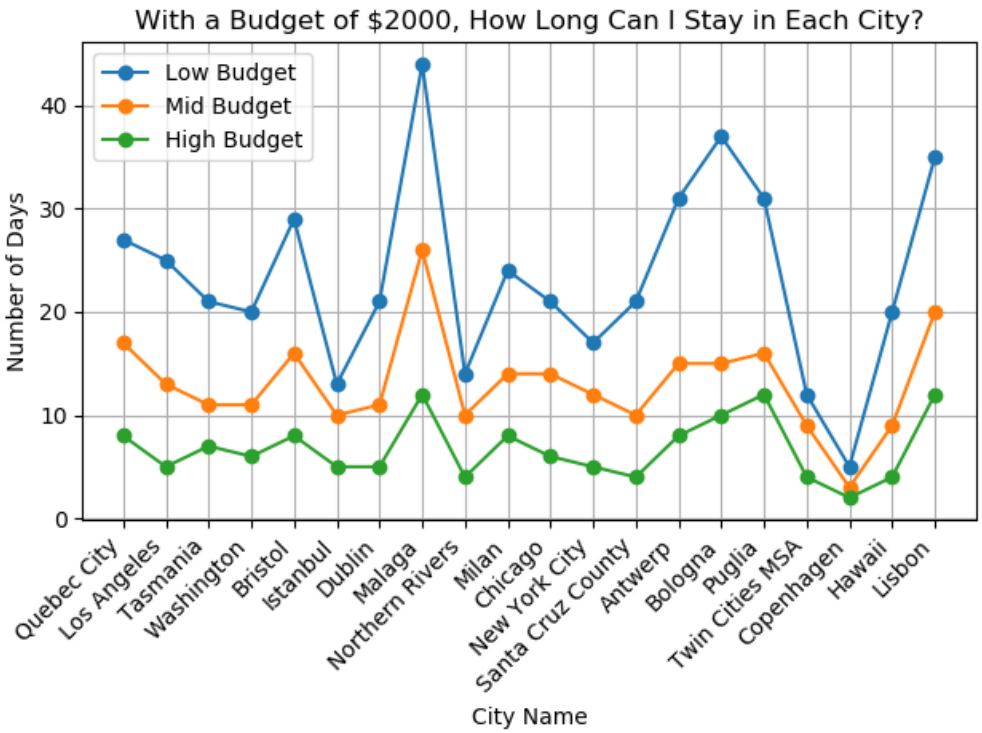

Figure 1: Comparing the Amount of Days You Can Spend in a City For Different Budget Levels

has the least amount of days you can spend in a city for the low, mid, and high budget, and Malaga has the highest amount of days you can spend in a city for the low, mid, and high budget. These occurrences agree with the cost-of-living index rankings.

5.2 Varying Budget Amounts

Continuing on, it would make sense if you had a larger budget, that you could spend more time in a place than if you had a smaller budget but spent the same amount of money each day in that city. Using `python`, we randomly selected 20 of our 89 cities to examine their different budget amounts as seen in Figure 2.

Looking at Figure 2, we can see that for any city, you can spend more days at the Mid level expenses with a budget of $5,000 than you could with a budget of $2,000 or $900. Additionally, this graph shows that with the same budget amount of $5,000, you can spend almost 6 times the number of days in Malaga than you would in Copenhagen. This relates back to the cost-of-living index ranking described in Section 5.1 in that Copenhagen is a more expensive city than Malaga.

6 Insights Gleaned

Before we even started the programming component of the project, we had troubles finding a dataset that would give us the information needed to plan vacations. Once we found the AirBnb dataset, we stopped looking for other data sources, but we should have kept looking for better sources. This dataset had lots of issues and if we would have examined the data more closely sooner, we could have had time to change our main dataset and reformulate our question. The Airbnb dataset was compromised mainly of American and European cities. Our original intention was to use data from all around the world, but we had to limit our cost-of-living source to accommodate this limited city

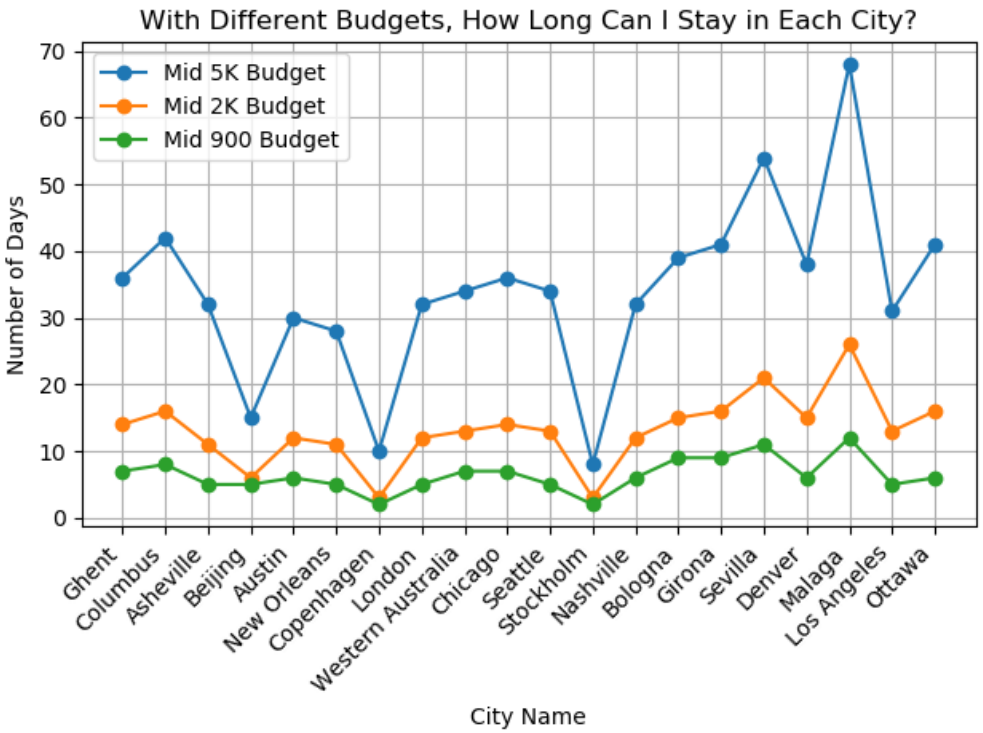

Figure 2: Comparing the Amount of Days You Can Spend in a City For Different Budget Amounts

selection. This dataset was based off of AirBnb Host input, so some of the features we needed from this dataset were misspelt or inaccurate. To combat the name misspellings, we used the name of the file, which was the name of the city, as a feature in our dataframe. With this additional features, houses from the same city were all properly grouped together, instead of having a plethora of groups for the same location. To combat the inaccurate input data, we should have implemented a threshold associated with each budget level average so that if a value was too far above or below this threshold, we would remove that entry and recalculate the average. There were some cities we expected to be more expensive or cheaper than what our program outputted. After examining our AirBnb housing data, we saw that some rooms were on the far end of the spectrum in what we would classify as anomalies and values that should be discarded from the data.

Originally we wanted our user to input different factors like what their eating habits are, or what type of transportation they want to use and we would formulate a travel plan for them based on these inputs. As we thought about it more, the purpose for our program was to make the travel planning process easier, and there is nothing easy about having to make decisions about everything involved in your trip. This is why we created different budget levels so we could create assumptions about what would be in each budget level and then users would be provided with a variety of options and they could choose the budget levels that fit their habits and their budget.

7 Transformation of the Problem Space in the Future

Since personal computers hit consumers in the 1970s, and the World Wide Web in the 1990s, the world of technology has exploded with innovations and incremental developments. The pace of change has become so increasingly fast that, in some areas, it can almost feel difficult to keep up and keep track of the current trends. It no longer suffices to design products and services for the present; they will soon be outdated and useless. Instead, much thought needs to go into the design

of these products and services to ensure that they can keep up with the ever-changing technological and social landscapes.

The first and most obvious problem space that can shift is common services used in travelling. An example of this would be how the lodging industry was dominated for years by hotels/motels, but recently has become a split market due to the rise of Airbnb use for leisure travelling. Hotels are still the standard for business travel, as companies are more comfortable compensating these expenses, and the hotel industry is dominant in city hubs where most business exchanges take place. However, many individuals have recently found it preferable to use Airbnb's service as a result of cost-efficiency, personalized experiences, app support, and much wider coverage where the hotel industry does not extend [10]. In the future, it would not be surprising if even more community-driven services arose in the lodging industry.

A more seemingly dry moving problem space to consider is that of global political and economic climates. These landscapes can significantly affect travelling internationally. For example, fluctuating currency exchange rates may greatly increase trip cost for travellers who wish to purchase goods locally. In more extreme examples, global crises may arise such as natural disasters or war that may make it unviable to travel or obtain permission to enter foreign borders. In these cases, destinations under these critical conditions should always be excluded from the list of suggestions to the user. Furthermore, regular seasonal economic fluctuations may also incur a "premium" that the traveller has to pay for goods and services due to high tourist demand.

Lastly, the evolution of leisure activities can be a major deciding factor of tourist activity. From the 1850s to the 1950s, mining and resource-based industries in the Rocky Mountains produced the majority of the revenue for communities; now, tourism generates much of the revenue from the attraction to new snowsports activities like skiing and snowboarding, as well as summer activities like hiking and mountain-biking [11].

All the ever-changing problem spaces can make it difficult to keep up with an individual's needs for a vacation. Trends would have to be continuously caught up with, and new services' data integrated into our suggestion app. Additionally, political climates and economic changes must be taken into account, in order to prevent suggesting a destination that could be too much overhead or even dangerous to the user. Finally, a way to filter destinations by vacation intent (Does the user just want to relax and see sights? Do they want to ski?) should be implemented to avoid flooding the user with unwanted suggestions.

8 Conclusions

After reaching a solid completion milestone for this project, we have concluded with several aspects regarding the effectiveness of our final result to address the initial problem at hand. Remember, the initial problem was that there are too many unknowns in trip planning, making it tedious to effectively compare two or more possible destinations against each other. In addition to driving up the pre-vacation period stress of the traveller, destinations of good fit are often overlooked in the planning process. However, even without running any experiments where users who plan normally are compared against users of our application, we are able to make a couple assertions about its utility.

The first assertion we are able to make about our application is that it significantly decreases the amount of time spent initially searching for a list of possible destinations. With a budget constraint, many of these destinations are ruled out. The destinations that are ruled out could have been a time-sink for the user, as the might have been intrigued, spent time researching them, only to be disappointed that they were, in fact, out of their budget. Instead, they can input a budget, and our application will output a list of reasonable places for them to visit in only a few seconds.

Next, we are able to assert that, by narrowing the list of reasonable places to visit down to a smaller subset, the application actually increases the users awareness of destinations they may have not thought previously about visiting. To give an example of this, suppose a traveller were to search on Google for "places to vacation". Initially, they might be overwhelmed by the sheer amount of places to visit that are being suggested by different sub-services of Google and third parties. Additionally, these places most likely include "hot spots" like Florence, Italy, but may completely omit similar and more affordable places like Bergamo, Italy. Thus, it is apparent that by suggesting a list of all places by budget, our application is able to make the user aware of locations that are a better fit to them than the "general" popular suggestions.

Lastly and most obviously, our application lets the user have a solid idea of what they can expect to spend on a vacation. This feature undoubtedly gives the potential traveller peace of mind with their final decision, and does not leave them up to guessing how much the vacation could hurt financially. Especially as our project takes on larger, newer datasets (i.e. global airfares) and is able to filter even more options for the user, we are confident it will be even more effective at improving a traveller's pre-vacation period.

References

[1] J. de Bloom, S. A. Geurts, T. W. Taris, S. Sonnentag, C. de Weerth, and M. A. Kompier, "Effects of vacation from work on health and well-being: Lots of fun, quickly gone," *Work & Stress*, vol. 24, no. 2, pp. 196–216, 2010.

[2] Z. Jing and Y. Fan, "Daily travel behavior and emotional well-being: Effects of trip mode, duration, purpose, and companionship," Sep 2018.

[3] J. Nawijn and J. de Bloom, "Pre-vacation time: Blessing or burden?," *Leisure Sciences*, vol. 35, pp. 33–44, 2013.

[4] W. T. Organization, "Unwto tourism highlights 2018 edition."

[5] M. Cox, "Inside airbnb," 2016.

[6] "About." https://www.kayak.com/about.

[7] "Fast facts." https://press.airbnb.com/fast-facts/.

[8] K. Amadeo, "How to compare the cost of living around the world," Feb 2019.

[9] "Measuring the cost of living worldwide," Mar 2017.

[10] K. Gyodi, "Airbnb in european cities: Business as usual or true sharing economy?," *Journal of Cleaner Production*, vol. 221, pp. 536–551, 2019.

[11] W. Kendall, "A brief economic history of colorado," *Demography Section, Colorado Department of Local Affairs*, pp. 3–5, 2002.