DSCI 510 Final Project Report
Kevin O'Kane

A Data Scientist's Ideal Home

Introduction
This project aims to determine which major cities in the United States are best for Data Scientists to call home. The project looks at several factors, including salaries, rent prices, and travel accessibility across 10 major U.S. cities including Austin, Boston, Chicago, Denver, Los Angeles, New York, San Diego, San Francisco, Seattle, and Washington DC.

Project Motivation
The motivation for this project was to provide a basic understanding for which city could provide the best place to live for entry level and early career Data Scientists in the United States. Upon completion of a Data Science master's degree program, most students will hope to have some optionality for their career across geographies and industries. This report aims to provide context for which regions could lead to the highest earning potential for data scientists. As someone who has lived in New York City for the last three years, it is a frequent occurrence to hear complaints about cost of living or rent prices in major cities in the United States. However, these complaints are rarely backed by data, rather anecdotes. Thus, this report aims to provide a quantitative backing to these discussions, specifically for data science professionals.

Description of Data Sources
This report considers three data sources to measure the average salaries of data science professionals across geographies, the average rent costs and tax burdens across the same geographies, and flight data to determine the distance from non-focus cities to these hubs of industry.

BuiltIn – To collect data for data science salaries, the popular tech job-listing website BuiltIn was used to find a plethora of job listings across these ten geographies. Through web scraping techniques, 480 job listings were found across the cities (48 per city). For each listing, the company, job title, location, company FTE, salary, and experience requirements were collected. It should be noted that some listings were missing information for FTE, salary, and/or experience. The search was filtered by 25 miles closest to each location, with jobs focused specifically between entry level and mid-tier experience (0-5 years of experience). The keyword used for the search was "Data Science." All levels of remote/hybrid/in-office positions were included.

Link: https://builtin.com

Apartments/NerdWallet – The second data source used included information from both the rental listing website apartments.com and the financial literacy platform NerdWallet. These sources combined to provide a brief picture at the costs of living in each location, to counterbalance the income from the first data source. The Apartments.com rental reports for each location were used to find average rent over the last year for studio and one-bedroom apartments. NerdWallet was used to find the approximate federal and state tax burden for a salary of $175,000, an approximation made based on the BuiltIn data. After coming up with both of these estimates, a calculation was made to determine the approximate "disposable income" that each city could provide, by subtracting the costs of rent and taxes from income. This data source proved difficult to web scrape from, as the dynamic nature specifically of most apartment listing websites led to difficulty in getting substantial data across locations.

Link: https://www.apartments.com/rent-market-trends/san-diego-ca/
Link: https://www.nerdwallet.com/article/taxes/california-state-tax

Amadeus – The third data source used an API from Amadeus, a travel information company. Two APIs were used to find information about flight paths from an inputted location to the 10 locations of interest. Through Amadeus, data regarding direct flight routes from airports of note to an inputted location was pulled. Additionally, latitude and longitude coordinates were pulled for all airports of interest. These coordinates were used to derive approximate flight time, by calculating a haversine distance, which accounts for the curvature of the Earth in taking distance between latitudes and longitudes, and dividing it by the average speed of a modern

commercial aircraft. The two APIs used for this project were the Airport Routes API, which was used to determine which cities had direct flights between each other, and the Airport and City Search API, which was used to determine the latitude and longitude of each airport. The APIs were fairly easy to use, and Amadeus provided examples through its GitHub + documentation on how to deploy these APIs. The biggest difficulty came in trying to determine how to calculate flight duration between locations, as APIs were limited in this area.

Link: https://developers.amadeus.com/self-service/category/flights/api-doc/airport-routes
Link: https://developers.amadeus.com/self-service/category/flights/api-doc/airport-and-city-search

Integrated Data Model
Below is a diagram of the datasets for this report. It's important to note that the calling of the APIs from Amadeus is dependent on the imputed airport in the provided Jupyter Notebook. More discussion in the following section.

jobs.csv (BuiltIn)	
location_id	INT
titles	STR
companies	STR
employees	INT
salaries	FLOAT
requirements	STR (Low or Mid)

<----------->

apartments.csv (NerdWallet/Apartments.com)	
location_id	INT
apartment type	STR (Studio or One Bedroom)
average rent	INT
average square footage	INT
approximate income taxes	INT

airport DataFrame – depedent on input (Amadeus)	
city	STR
geoCode	DICT
latitude	FLOAT
longitude	FLOAT
Flight Duration	FLOAT

Analysis of Data Visualizations
The report provides four data visualizations to demonstrate the findings of the data. The first three aim to address which cities are cheapest to live in and pay the most. The first shows a confidence interval for salaries across entry and mid-level positions in each city. The second shows the average rent and square footage for studio and one-bedroom apartments in each city. The third demonstrates a basic and approximate disposable income of data scientists across these cities, which is calculated by taking the difference between salary and rent/tax burden. Visualizations can be found in the provided notebook (final_report.ipynb).

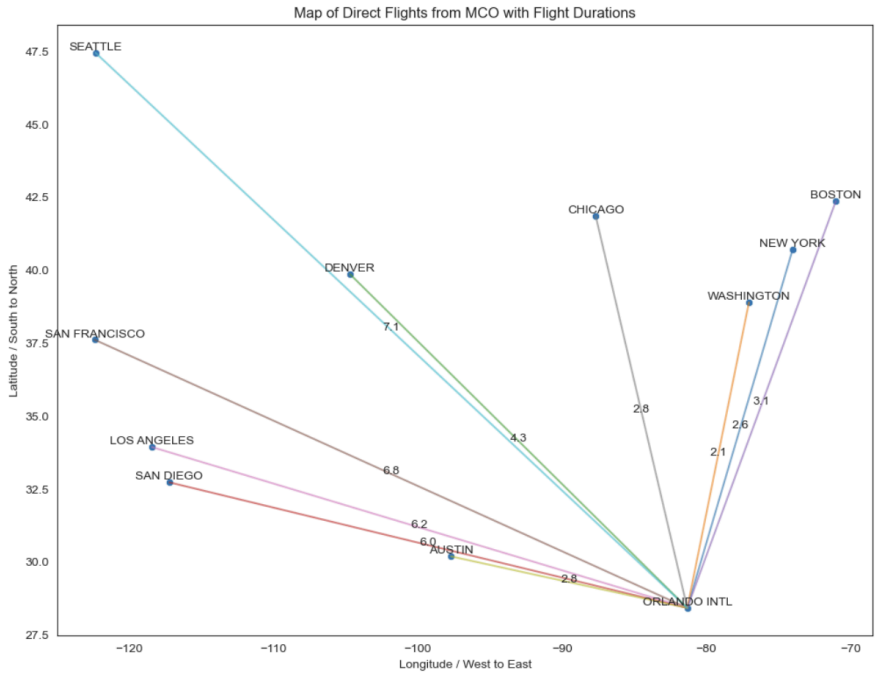
Figure 1: Salaries by Location and Experience Level – Figure 1 shows boxplots for salaries per location and experience level. From the diagram, we can conclude several takeaways across the focus geographies. Overall, there does not appear to be any significant outliers in either direction on the salary front, though there are differences. Cities like Denver, Washington DC, and San Diego show lower salaries than the other cities. Boston demonstrates the greatest value for experience, as the difference between its pay at the entry level and mid level is substantial. Cities like New York City, San Francisco, and Seattle pay well for entry level workers.

Figure 2: Apartments Cost and Size by City and Arrangement – Figure 2 demonstrates the approximate rent prices and average square footage for apartments in the focus cities. Notably Boston and New York City are the most expensive cities to live in, with New York also providing below average square footage. Cities like Denver and Austin provide both cheaper rent and larger apartments, which make them enticing despite lower salaries demonstrated in Figure 1. San Francisco is both expensive and provides the smallest apartments via square feet. Most cities tend to cluster for both rent and square footage, but it is noticeable that cities have a larger difference in square feet for one-bedroom apartments over studios.

Figure 3: Income after Taxes and Rent by City – Figure 3 displays a bar graph sorted by "disposable income," or the amount of cash remaining after both state and federal income taxes and annual rent charges. Seattle and Austin seem to be the big winners for these cities, as they are the clear two highest for this measure. This makes sense because both Texas and Washington do not charge residents income taxes. New York City is the lowest for disposable income, which is likely due to the high rent charges for living in the Big Apple. Washington DC has the highest "state" income tax at roughly 14%, though it is notable that given it is not within a state, city taxes were included in the calculation in its place. In terms of state tax, California has the highest behind DC. However, it seems that the difference in income over the other cities in the study proves to level out the take-home pay for Californians despite high tax rates.

Figure 4: Map of Direct Flights from Inputted Airports – The fourth figure displays less quantitative data than the previous figures. The goal of this figure is to answer the question: "How far is home?" The map requires an inputted United States airport code. Once imputed, the APIs are called to produce a scatter plot of latitude and longitude coordinates, with lines acting as flight paths to the cities of interest in this study. Lines will only appear from the origin/inputted airport if there are direct flights from the airport to the focus cities. Each line also demonstrates the approximate flight duration from the origin to the geographies of interest for this study. Again, the idea for this visualization is to provide an understanding of how far away each city is from a specific city, and whether or not that city is reachable via direct flight.

An example of the scatter plot produced is shown below for the origin airport code MCO, which is the International Airport of Orlando, FL. As the map shows, each city of interest is accessible via direct flight from Orlando, with flight durations between just over 2 hours for Washington DC and over 7 hours for Seattle.



Conclusions and Future Work

Overall, the report produces interesting takeaways regarding salaries and cost of living for data scientists across the 10 focus geographies. The main takeaways seem to be that salaries are largely similar across these geographies, and that tax burden and rental costs will have a more significant impact on disposable income. The map included in Figure 4 also adds a more qualitative factor into decision making, depending on the distance to towns that data scientists might have to visit frequently (i.e. hometowns, other family locations, business trips, etc.). If given more time to continue this report, additional information for cost differences would add to understanding of disposable income in these geographies. For example, while New York City had the highest cost of living, it is the only city in the study with the public transportation infrastructure to not require a car, leading to savings in other non-rent areas. Also, potential differences in food costs could be analyzed to see if there's substantial differences, though it likely would not show significant differences across locations. Overall, a career in data science seems a fruitful path across geographies for any and all analytically-minded students.