IST 652 - Scripting for Data Analysis
Professor Debbie Landoswki
Team members: Nora Lin and Michael Morrey

Final Project Report

Introduction:

In today's era of big data, roles relating to data science are becoming increasingly important and in demand. Careers in data science are expected to grow rapidly over the next decade. Data science is thought of as a broad term, and data scientist roles can be found in a wide range of disciplines. Data scientist roles can also include many titles, including Data Scientist, Machine Learning Specialist, Data Analyst, and Data Engineer.

Glassdoor.com is a website that provides insight into companies and roles within companies. Employees and interviewees have the ability to anonymously submit company ratings and reviews, salaries, and detailed descriptions of specific roles, including data scientist roles. For people seeking careers in the highly sought after field of data science, careful analysis of the data found in Glassdoor.com can provide insight into finding the most optimal roles specific to each individual.

Zillow.com is another website that provides insight into housing prices. Historical data is tracked on Zillow, providing a means to observe the volatility variation in each market. The combination of glassdoor and zillow can provide some insight into which jobs have salaries that can afford the housing prices by location.

About the data:

The Glassdoor data set being evaluated comes from a CSV file found on Kaggle.com, and contains scrapped job postings from Glassdoor.com relating to the job of "data scientist". The original data set can be found at the following link:
https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor

The raw dataset contains 42 columns and 742 rows. The following table displays the name and description of each column, as well as an example of data that can be found within each column.

Fields	Description	Example
index	Index	1
Job Title	The title of the job, e.g. Data Scientist, Junior Data Scientist, etc.	Data Scientist

IST 652 - Scripting for Data Analysis
Professor Debbie Landoswki
Team members: Nora Lin and Michael Morrey

Salary Estimate	Range of Salary and the source	$53k - $91k (Glassdoor est.)
Job Description	Indicates what qualities that company wants and what is expected out of the job title	Data Scientist Location: Albuquerque, NM Education Required: Bachelor's degree required, preferably in math, engineering, business, or the sciences. Skills Required: Bachelor's Degree in relevant field, e.g
Rating	Rating of the company	3.8
Company Name	Name of the company	Tecolote Research 3.8
Location	Location of the job (city, state)	Albuquerque, NM
Headquarters	Location of the headquarters of the company (city, state)	Goleta, CA
Size	Range of the number of employees working in the company	501-1000
Founded	Company founded in year	1973
Type of ownership	Type of ownership	Company - Private
Industry	Industry of the company	Aerospace & Defense
Sector	Sector of the company	Aerospace & Defense
Revenue	Revenue of the company	$50to $100 million (USD)
Competitors	Company competitors. If not successfully scrubbed, then -1	-1
Hourly	If the job is paid hourly (0=no, 1=yes)	0
Employer Provided	If the employer name is provided (0=no, 1=yes)	0
Lower salary	Lower end of the salary range	53
Upper Salary	Higher end of the salary range	91
Avg Salary(K)	Average salary of the job posting	72

company_txt	Name of the company	Tecolote Research
Job Location	State abbreviation of the job location	NM
Age	Age of the company in years	48
Python	Python programming required (0=no, 1=yes)	1
Spark	Spark programming required (0=no, 1=yes)	0
AWS	AWS programming required (0=no, 1=yes)	0
Excel	Excel required (0=no, 1=yes)	1
SQL	SQL required (0=no, 1=yes)	0
SAS	SAS required (0=no, 1=yes)	1
Pytorch	Pytorch required (0=no, 1=yes)	0
Scikit	Scikit required (0=no, 1=yes)	0
Tensor	Tensor required (0=no, 1=yes)	0
Hadoop	Hadoop required (0=no, 1=yes)	0
Tableau	Tableau required (0=no, 1=yes)	1
Bi	BI required (0=no, 1=yes)	0
Flink	Flink required (0=no, 1=yes)	0
Mongo	Mongo required (0=no, 1=yes)	0
Google_An	Google analytics required (0=no, 1=yes)	0
Job_title_sim	Simplification of the job title (ex. Data scientist, analyst)	Data scientist

seniority_by_title	Seniority by title (sr, jr, na)	na
Degree	Degree level required (M=Masters, P=PhD, na=not specified)	M

The Zillow dataset can also be found on Zillow's website. It has 29,770 rows and 35 columns. The original dataset can be found here: http://files.zillowstatic.com/research/public_csvs/zhvi/Zip_zhvi_uc_sfr_month.csv
The columns are as follow:

Fields	Description	Example
RegionID	index	61639
SizeRank	Size rank of city	0
RegionName	Size code	10025
RegionType	Type of region	Zip
StateName	State Name	NY
State	State	NY
City	City	New York
Metro	Metropolitan area	New York- Newark - Jersey City
CountyName	County name	New York County
1996-01-31 to 2022-04-30	Housing value for each date	2.9 e06

Preprocessing:

To begin preprocessing the data, the pandas, mumpy, matplotplib, and seaborn packages were imported. These were the necessary packages that would be used for analysis. After importing the Glassdoor CSV file and converting it to a data frame, the contents of the data frame were examined. Upon reviewing the contents, it was apparent that several columns would not be needed for analysis. The following columns were removed from the data set.

- **Index** - Not useful information
- **Job Title** - More simplified job titles found in job_title_sim column
- **Salary Estimate** - More useful salary ranges found in Lower Salary, Upper Salary, and Avg Salary(K) columns
- **Job Description** - Long text string that would not be useful for our purposes
- **Headquarters -** Not relevant for our purposes
- **Size** - Not relevant for our purposes

- **Founded -** Not relevant for our purposes
- **Type of ownership** - Not relevant for our purposes
- **Revenue -** Not relevant for our purposes
- **Competitors** - Blank and multivariate data
- **Hourly** - Very few jobs were hourly
- **Employer provided** - Not useful information
- **Company Name -** Cleaner values found in company_txt column
- **Age -** Not relevant for our purposes
- **Seniority_by_title -** Lots of missing data

After narrowing down the data frame to the desired columns, many of the remaining column names were renamed for consistency and clarity. The columns were also slightly rearranged to bring Job_Title and Company_Name to the beginning of the data frame. This allowed for the column grouping to make more logical sense.

Basic summary statistics were then run on the numerical data, as shown in the following snippet. This summary revealed that the Rating column had a minimum value of -1, which meant that those values were not successfully scrubbed. Those values were then dropped. Each categorical column was then evaluated to determine if -1 was present, and it was revealed that the Industry and Sector columns contained -1 values. These values were replaced with the value of "Unknown".

	Rating	Lower_Salary	Upper_Salary	Avg_Salary	Python	Spark
count	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000
mean	3.618868	74.754717	128.214286	101.484501	0.528302	0.225067
std	0.801210	30.945892	45.128650	37.482449	0.499535	0.417908
min	-1.000000	15.000000	16.000000	15.500000	0.000000	0.000000
25%	3.300000	52.000000	96.000000	73.500000	0.000000	0.000000
50%	3.700000	69.500000	124.000000	97.500000	1.000000	0.000000
75%	4.000000	91.000000	155.000000	122.500000	1.000000	0.000000
max	5.000000	202.000000	306.000000	254.000000	1.000000	1.000000

The Kaggle dataset description mentioned "na" values being present in the Degree column, indicating that a Masters degree or PhD were not required for the job. To allow for more clarity, these "na" values were renamed to "Not Required".

The last step taken was to add a column called "Total_Skills", indicating the total number of skills required for the job. It was calculated by adding the skills columns together, which were indicated with a 1 for required and 0 for not required.

After importing the Zillow CSV file and converting it to a data frame, the contents of the data frame were examined. Several columns were dropped to maintain the scope of this assignment. All date columns were removed except for dates pertaining to 2021. This was done to maintain consistency in time periods between the glassdoor data. The pandas function pd.melt() was used to transform the data so that each date is a row instead of a column. State and City columns were combined to create a new column titled "Location". Grouping by location,

the median housing price was calculated. The final dataset consists of locations in the format of (City, State) and the average housing value.

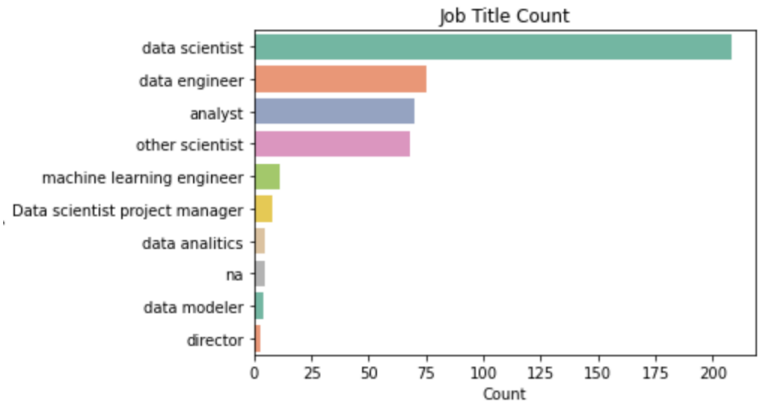
Lastly, a final dataset was created by merging the Glassdoor and Zillow dataset. There are 434 rows and 29 columns. Only 23 rows were dropped during the merge. Following the merge, a new column titled "Monthly_Income" was added to the dataset, which calculates a monthly income based on the average annual salary. At that point, the dataset was fully cleaned and ready to be analyzed.

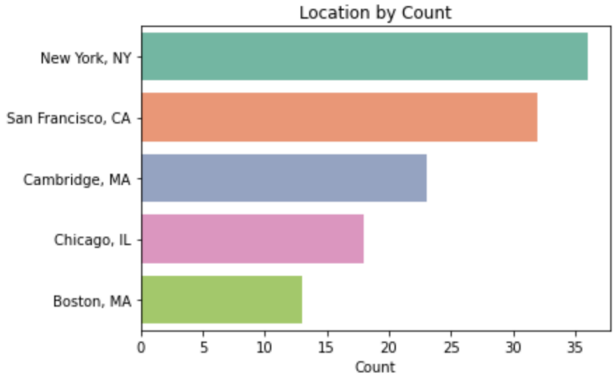
Method of Analysis:

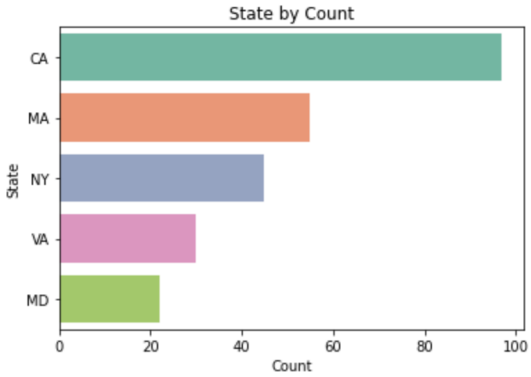
Initial data exploration was conducted on both datasets. From the Glassdoor dataset, we wanted to analyze and answer the following questions with our exploratory data:

1. What is the distribution of job titles?
2. How are the average ratings distributed among the job titles?
3. What are the top five cities that have the most jobs?
4. What are the top five states that have the most jobs? Do they match what the states from the top five cities?
5. What is the distribution of salaries by job title?
6. Do certain job titles require specific degrees?
7. Which company has the highest average salary among data scientists?
8. Which company has the lowest average salary among data scientists?
9. What are the most highly sought after data science skills for each job title?

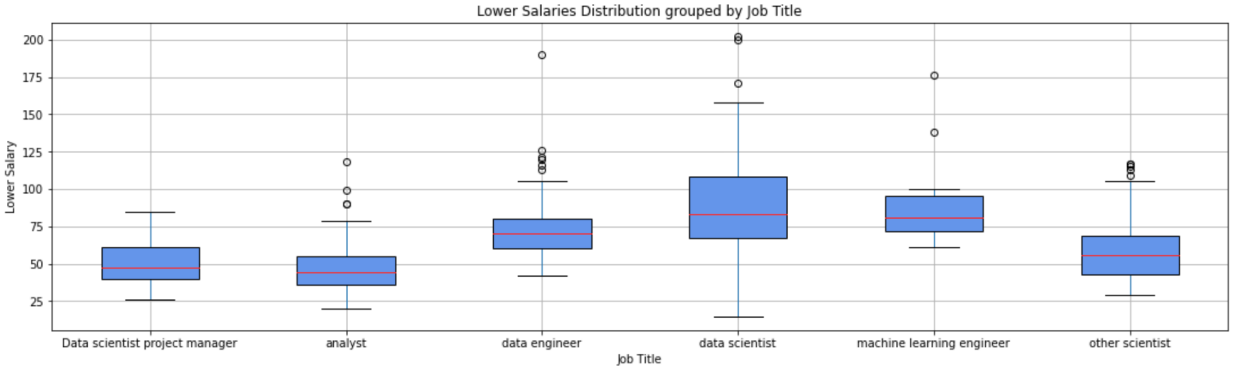
The following graph shows the count of job titles across the dataset. "Data scientist" has the highest count while "director" has the lowest count. For the purpose of this assignment, the following job titles were dropped from the dataset due to low counts: "director", "data modeler", "na", and "data analytics". Analysis was also performed on the rating of the jobs based on job titles. "Data engineers" and "Data scientist project manager" were tied for the highest rating at 3.8. "Data scientist" had the second highest rating at 3.79, followed by "analyst" at 3.65, "machine learning engineer" at 3.5, and "other scientist" at 3.4. The following figures show the top 5 locations and states by count of jobs. New York, NY has the highest count of jobs followed by San Francisco, CA. These are not only two major cities, but also two major tech cities. By state, California and Massachuseets have the highest number of jobs.



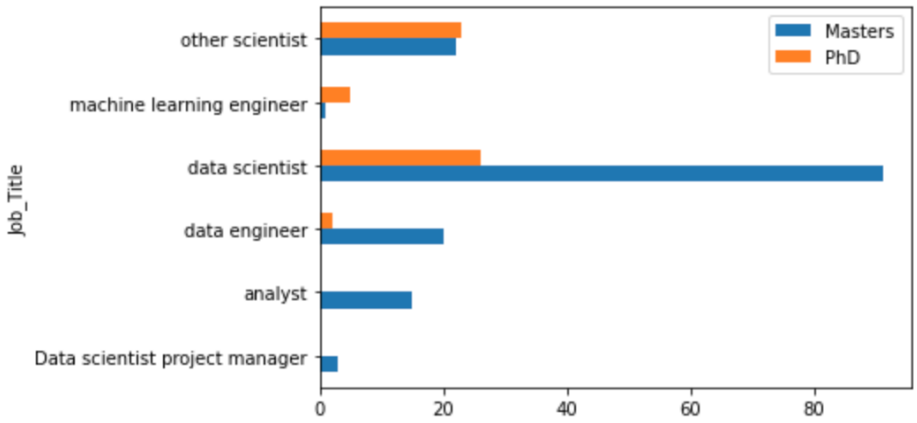




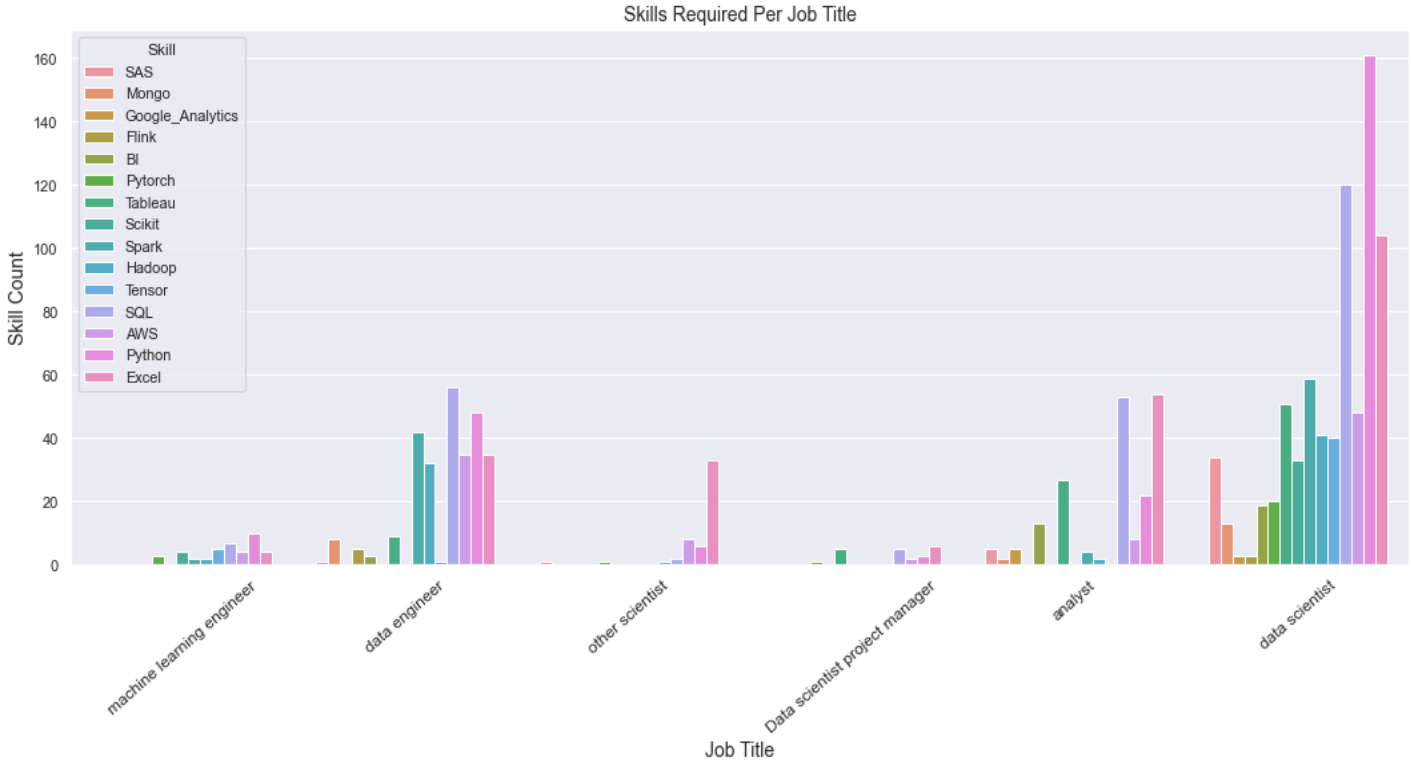
The following boxplot figures examine the upper salaries and lower salaries by job title. “Data engineers”, “analysts” and “machine learning engineers” have smaller variations in upper salaries compared to “data scientists”. There are some outliers within “data scientist” jobs that range above $200,000. Most of the jobs are not skewed except for “machine learning engineer” and “data scientist project manager”. Again, we can see that data engineers, analysts, and machine learning engineers have smaller variations in upper salaries compared to data scientists.



From the glassdoor dataset we also examined if certain jobs require different degrees. The following horizontal barchart shows that across the board, Masters degrees are required. However, there are only a couple of job titles that require PhD degrees.



Lastly, from the Glassdoor dataset we explored which data science skills were the most highly sought after for each job title. The following bar plot breaks down the total count of required skills in the job listings for each job title. For data scientists, Python was listed as required the most frequently by a considerable margin. The next highest skills were SQL and Excel, respectively. Python was also listed the most frequently for machine learning engineers. Data engineer and analyst positions both required a heavy amount of SQL skills. Analysts appear to show Excel being required the most frequently with a very slight edge over SQL. Excel appears to be relatively high compared to other skills for every job title.

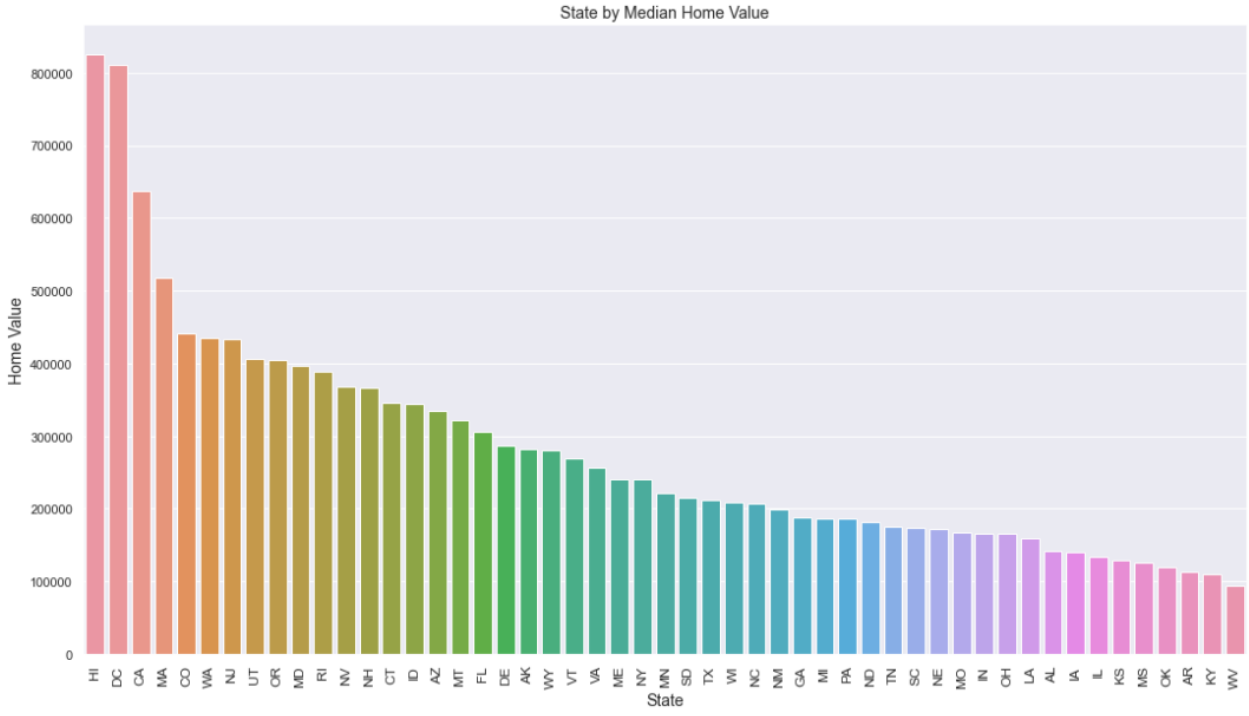


IST 652 - Scripting for Data Analysis
Professor Debbie Landoswki
Team members: Nora Lin and Michael Morrey

Following initial data exploration on the Glassdoor dataset, the Zillow dataset was then similarly analyzed to answer the following questions:

1. How do the states compare to each other in median home value?
2. What are the top five cities in median home value?
3. What are the bottom five cities in median home value?
4. What is the distribution of home values by state?

To understand how the states compare to each other in median home value, the following bar plot figure was created. The plot compares all 50 states plus Washington D.C. and their median home values, sorting the home values in descending order. Hawaii and Washington D.C. were the top two states in median home value by far, both with median prices over $800,000. California and Massachusetts followed, with median home values of over $600,000 and $500,000, respectively. Conversely, West Virginia had the lowest median value with less than $100,000. Kentucky and Arkansas finished slightly above West Virginia, with values of just over $100,000.



Following the analysis on the states, the top 10 and bottom 10 cities in median home value were evaluated, as shown in the two graphics on the following page. The “Top 10 Cities Median Home Value” graphic reveals three cities in California and two cities in New York that rank in the top 10. The top city of Atherton, California had a whopping median value of

$8,301,180.34. One surprising city that was found in the top 10 was Teton Village, Wyoming, with a median value of $4,343,340.86.

From the "Bottom 10 Cities Median Home Value" graphic, we can see that seven out of the ten cities were found in West Virginia. The city with the lowest median home value was Gary, West Virginia, with the median price of $16,446.42. Of the three remaining cities, two were Pennsylvania cities, and the other was found in Illinois.

Top 10 Cities Median Home Value

Location	Value
Atherton, CA	8301180.34
Palm Beach, FL	8092794.82
Sagaponack, NY	6247408.65
Hollandale, MS	5464304.59
Malibu, CA	4802395.75
Water Mill, NY	4515366.51
Aspen, CO	4351876.76
Teton Village, WY	4343340.86
Medina, WA	4333109.55
Montecito, CA	4276887.61

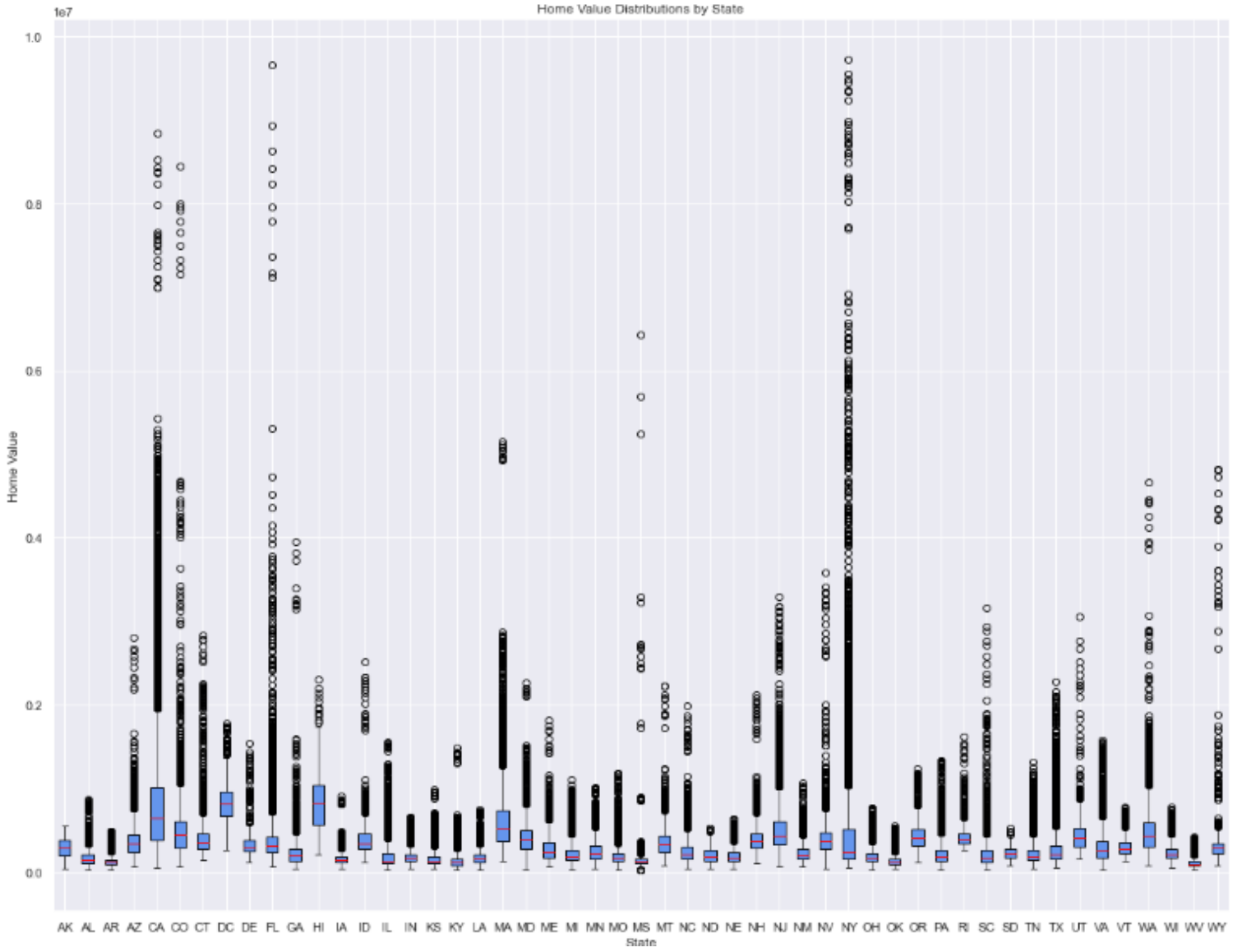
Bottom 10 Cities Median Home Value

Location	Value
Gary, WV	16446.42
Elkhorn, WV	19520.52
War, WV	20561.27
Berwind, WV	21173.08
Girardville, PA	22318.31
Cairo, IL	22627.62
Mahanoy City, PA	22722.46
Keystone, WV	22752.20
Raysal, WV	23161.30
Rhodell, WV	24713.58

The last method of exploratory analysis on the Zillow dataset was done by analyzing the distributions of home values among each state, as shown in the box plot figure on the following page. The first visual that jumps out is the large number of outliers in certain states. New York, in particular, appears to have the largest variance in outliers, indicating a large number of very expensive homes. The most expensive home in New York approaches $10,000,000. Other noticeable states with high variance in outliers include California, Florida, and Colorado. Conversely, Alaska appears to be the only state with no outliers.

Another item of note is the variance within the interquartile range (25th percentile to 75th percentile). Within this range, California, Hawaii, Massachusetts, New York, and Washington D.C. appear to have the highest variance in home prices. Conversely, states with the lowest variance within the interquartile range tend to fall in the midwest and southern regions of the country. These states include Arkansas, West Virginia, Oklahoma, Mississippi, Iowa, Indiana, Kansas, Kentucky, and Louisiana. This means that if someone were to look for homes in these states, they would likely find very similar home prices for the most part, with some outliers.

The last portion from the box plox below that was analyzed was the minimum home values across the states. Many of the states appear to have minimum home values at a very similar level. However, some states jump out as having higher minimum home values. These states include Hawaii, Washington D.C., Massachusetts, and Rhode Island. Based on this data, people looking to live in the cheapest areas may want to stay away from these states.



Program descriptions:

Program 1: The main question we wanted to answer with program 1 was can we create a custom filter based on user input similar to the filter effect on the Glassdoor website? The program first obtains several inputs from the user based on the following questions:

- "Please enter the job title from the following options: data scientist project manager, analyst, data engineer, data scientist, machine learning engineer"
- "Please enter the lowest acceptable job rating out of 5: "
- "Please enter the highest acceptable job rating out of 5: "
- "Please enter your desired location in the format of City, State: "
- "Please enter the lower bound of your desired average salary: "
- "Please enter the highest bound of your desired averaged salary: "

The program then prints out the requirements that the user has inputted. It then uses the requirements to create a new dataframe. If there are no rows that match the user requirements, the program outputs “Sorry we could not find any results! Try modifying your search”. If the program finds only one job that matches the results, then the program prints out the name of the company. If the program finds more than one job, then the program prints out the names of the companies.

The following image displays the output of the first program after entering the required inputs. The inputs we used are also displayed following each input question. From these inputs, the program found six Glassdoor job entries that fit the requirements, and listed “Biz2Credit Inc” and “Farportal” as the top two company names.

```
Please enter your desired job title from the following: Data scientist project manager,analyst,data engineer, data scientis
t, machine learning engineer data scientist
Please enter your lowest acceptable job rating out of 5: 3.2
Please enter your highest acceptable job rating out of 5: 4
Please enter your desired location in the format of City, State (abbreviated): New York, NY
Please enter the lower bound of your desired average salary (example: 50 = $50,000): 50
Please enter the upper bound of your desired average salary (example: 100 = $100,000): 100
----
Your requirements were: data scientist with a job rating between 3.2 and 4.0 located in New York, NY and salary between 50.0
and 100.0
----
From our Glassdoor database, it looks like we found 6 job entries that fulfills your requirements!
The top two company names are: Biz2Credit Inc and Fareportal
```

Additionally, all six job entries that fit the requirements were outputted to the following data frame.

Job_Title	Company_Name	Rating	Location	Industry	Sector	Lower_Salary	Upper_Salary	Avg_Salary	Job_State	...	Tensor	Hadoop	Tableau
data scientist	PA Consulting	3.4	New York, NY	Consulting	Business Services	71	123	97.0	NY	...	0	1	1
data scientist	Fareportal	3.8	New York, NY	Travel Agencies	Travel & Tourism	53	96	74.5	NY	...	0	0	0
data scientist	IHS Markit	3.5	New York, NY	Consulting	Business Services	61	100	80.5	NY	...	0	0	0
data scientist	Biz2Credit Inc	4.0	New York, NY	Lending	Finance	63	111	87.0	NY	...	0	0	0
data scientist	Strategic Financial Solutions	4.0	New York, NY	Consumer Product Rental	Consumer Services	71	124	97.5	NY	...	0	0	0
data scientist	Remedy BPCI Partners, LLC.	3.4	New York, NY	Health Care Services & Hospitals	Health Care	69	121	95.0	NY	...	0	0	1

These results show that through proper cleaning and analysis of the Glassdoor data set, streamlined job listing results can be obtained specific to each potential job candidate. This can allow for the most efficient job searches, eliminating the need for sifting through many jobs in a city that don’t fit the user’s requirements.

Program 2: The second question we wanted to answer was based on job salary and user input on mortgage information, can we identify jobs in affordable housing areas? The goal was to determine affordability based on the payment to income (PTI) ratio. The program begins by obtaining user input from the following questions:

- "Please enter the job title from the following options: data scientist project manager, analyst, data engineer, data scientist, machine learning engineer"
- "Please enter your maximum payment to income value (example: .25)"
- "Please enter the down payment amount you will pay (example: 20000)"
- "Please enter how many years your fixed mortgage will be (15 or 30)"
- "Please enter the interest rate you expect to pay (example: .05)"

The user inputs are then printed out. The program then calculates a monthly payment on the median home value in each city based on the home price, down payment, mortgage type (15 or 30 year fixed mortgage), and interest rate, and stores that value in a new column called "Monthly_Pmt". Another column is then created in the dataframe called "PTI", which is calculated by dividing the "Monthly_Pmt" column by the "Monthly_Income" column. That column value is then rounded to four decimal places. The data frame is then filtered out by pulling all rows where the "Job_Title" column is equal to the job title that was specified in the user input, and where the "PTI" column is less than the maximum PTI value that was specified in the user input. Those results are then sorted by PTI in descending order. The program then outputs the text "Based on your input values, the following CSV file provides the most affordable locations for you, ordered by PTI descending". The final steps are to output the final data frame into a CSV file titled "Affordable_Locations_Jobs.csv", and then to display the top 10 results of the data frame.

The following image displays the output of the first program after entering the required inputs. The inputs show that the user is looking for a "machine learning engineer" job title, have a maximum desired PTI of 30%, will pay a $30,000 down payment on their mortgage, obtain a 30 year fixed loan, and expects to pay 4% on interest. The results also indicate that a CSV file was exported, containing all of the affordable job locations based on the user inputs.

```
Please enter the job title from the following: Data scientist project manager,analyst,data engineer, data scientist, machine
learning engineer: machine learning engineer
Please enter your maximum payment to income value (example: .25): .3
Please enter the down payment amount you will pay (example: 20000): 30000
Please enter how many years your fixed mortgage wil be (15 or 30): 30
Please enter the interest rate you expect to pay (example: .05): .04
----
Your requirements were: machine learning engineer with a maximum PTI of 0.3, down payment of 30000.0, and an expected 0.04 i
nterest rate on a 30 year fixed mortgage.
----
Based on your input values, the following CSV file provides the most affordable locations for you, ordered by PTI descendin
g.
```

Additionally, the following image displays the CSV output in Excel format.

	Job_Title	Company_Name	Location	Industry	Avg_Salary	Value	Monthly_Income	Monthly_Pmt	PTI
337	machine learning engineer	Sage Intacct	San Francisco, CA	Computer Hardware & Software	232.5	1910648.25	19375	5432.98	0.2804
367	machine learning engineer	Mteq	Fort Belvoir, VA	Aerospace & Defense	87	674406.79	7250	1861.62	0.2568
391	machine learning engineer	Information Builders	New York, NY	Computer Hardware & Software	125	926612.7	10416.67	2590.21	0.2487
267	machine learning engineer	Cboe Global Markets	Lenexa, KS	Stock Exchanges	87	408074.53	7250	1092.22	0.1507
126	machine learning engineer	Stratagem Group	Aurora, CO	Aerospace & Defense	100.5	452608.79	8375	1220.87	0.1458
174	machine learning engineer	Tempus Labs	Chicago, IL	Biotech & Pharmaceuticals	133	391723.78	11083.33	1044.98	0.0943
432	machine learning engineer	Software Engineering Institute	Pittsburgh, PA	Colleges & Universities	107.5	226526.87	8958.33	567.74	0.0634
168	machine learning engineer	Software Engineering Institute	Pittsburgh, PA	Colleges & Universities	120	226526.87	10000	567.74	0.0568
185	machine learning engineer	Software Engineering Institute	Pittsburgh, PA	Colleges & Universities	124	226526.87	10333.33	567.74	0.0549

These results show that through properly cleaning and merging the Glassdoor and Zillow datasets, customized job listings can be obtained specific to each user who may be deciding what cities or states to live in based on affordability. This provides users a powerful tool for quickly identifying jobs in cities that meet their specific needs without having to calculate cost of living separately.

Conclusion:

From exploratory analysis on both the Glassdoor and Zillow datasets, several conclusions were drawn. From the Glassdoor dataset, we observed that the majority of the dataset consisted of “data scientist” job titles. Out of the job titles relating to “data scientist”, we concluded that the most highly rated jobs, from highest to lowest, are data engineer, data scientist, analyst, machine learning engineer, and other scientist. Additionally, when considering cities to live in for data science related positions, tech cities tend to have the most available positions. These cities include New York City, San Francisco, Cambridge, Chicago, and Boston. Similarly, when considering states to live in, California, Massachusetts, New York, Virginia, and Maryland have the most data science related positions. Three of five states from this list contain the cities that have the most available positions. From the distributions of salaries by job title, we concluded that the most highly paid job titles in the field are data scientist, machine learning engineer, and data engineer. If applying for any one of these three highly paid positions, there is a higher probability that either a Masters degree or PhD will be required. Lastly, we concluded that the most universally sought after skills for data science positions are Python, SQL, and Excel.

From the Zillow dataset, we concluded that Hawaii and Washington D.C. are by far the most expensive states to live in, in terms of median home value. After observing the top 10 and bottom 10 cities by median home value, we concluded that California and New York both contain cities that are very expensive to live in. Lastly, after examining the distribution of home values among each state, a person who is looking for a home in New York, California, Florida, and Colorado can expect to find the highest variance in home prices. Conversely, it can be concluded that states in the midwest and southern regions will likely have lower variance.

From the two programs that were developed, we can conclude that user input programs can be very valuable for both job-searching and housing market tools. The first program, which allows users to custom filter the Glassdoor dataset, outputs the company name or names

depending on the result matches. This program is similar to the filter option on Glassdoor's website. However, our program can provide more concise results. The second program, which allows users to custom filter the merged Zillow and Glassdoor dataset, provides affordability output that goes beyond what is offered on Zillow. This can greatly increase efficiency for potential job candidates and home buyers when considering which jobs to apply for, and in which locations to look for those jobs.

In conclusion, our project showed that the best paid data science-related titles were data scientist, machine learning engineer, and data engineer. All three of these jobs were likely to require advanced degrees, such as masters or PhD. The most expensive areas based on median housing value were Hawaii, Washington D.C, California and Massachusetts. Our user input programs were very valuable for job-search and housing market tools. The ability to streamline the process based on user inputs shows the power of these coding techniques.

Roles:

Michael:

- Obtained and cleaned glassdoor dataset
- Conducted exploratory analysis on zillow data set
- Wrote introduction and about the data
- Wrote method of analysis for zillow data
- Wrote program 2 description
- Wrote program 2 code
- Made half of the slides

Nora:

- Obtained and cleaned zillow dataset
- Conducted exploratory analysis on glassdoor dataset
- Wrote method of analysis for glassdoor data
- Wrote program 1 description
- Wrote program 1 code
- Made half of the slides