2022

DATA SCIENCE JOB POSTING DATA ANALYSIS IN SAS

HEMALATHA RAMAKRISHNA

CIS 5250: VISUAL ANALYTICS | SAS PROJECT

MS IN INFORMATION SYSTEMS | CALIFORNIA STATE UNIVERSITY, LOS ANGELES

Table of Contents

Introduction

The term "data science" and the practice itself has evolved over the years. Its popularity has grown considerably due to recent innovations in data collection, technology, and mass data production across the globe. The procedures to collect, analyze, and interpret data paved the way for data science to become a popular field today. With the flood of new information and businesses seeking new ways to increase profit and make better decisions, the data science job market expanded worldwide.

Data Science is a highly prominent emerging discipline and career domain. According to LinkedIn research, Data Science is the fastest-growing new employment category. Professionals with training in Data Science received a remarkable 37% hiring increase over the last three years. Data science is reshaping practically every business and gaining traction daily. The use of data analytics in the industry is something companies can no longer afford to ignore. Data science is sometimes complex and challenging to understand, but the businesses that implement systems and strategies to collect, analyze, and use data, will experience quantifiable benefits in numerous areas of their operation.

Over 60% of respondents to a 2015 CapGemini study agreed that failing to use big data could lead to irrelevance and loss of competitiveness. The same study described the willingness to utilize data and emerging technologies in digital transformation. Modern businesses must embrace newer technologies to communicate with and understand their customers. Digital transformation also relates to how businesses are run and data is collected. Automated data collection and analytics are necessary for a company’s digital transformation. [1]

In a day, 500 million tweets are sent, 294 billion emails are delivered, and 4 petabytes of data are created on Facebook. Finding the right skilled individuals to transform data into insights is essential for a successful business. The field's newness has made the data science job role an evolving title. Today's data scientists must possess the abilities to collect, clean, extract, transform and load data and must be able to communicate the findings in both written and spoken form.

Glassdoor listed data scientist as the number 1 career, but it wasn't just top of the list for tech. It topped every industry. The fast-paced growth of data science jobs has been met with a severe lack of qualified candidates. And businesses that do hire for the role of data science jobs often have no idea how to utilize their skills effectively. [2]

The objective of the study

In this project, we're looking at the data science job opportunity connected to their salary, location, firm, sector, industry type, and skills to determine whether there's a link between these factors. This dataset contains data science job roles and descriptions in different companies and approximate salaries offered for various positions in other companies. This data set contains all the information on data science job posting on Glassdoor, like company name, industry type, sector, minimum and maximum salary offered, location of the company, skills required, and job description of the role. This analysis will help us find the key factors driving the demand for data science jobs. We will be able to understand the inclination of organizations toward data science business strategies and the rising adoption of an advanced career in creating opportunities for the data science job through Glassdoor data analysis. [3]

Data Description

This data set is regarding Data Science Job Posting on Glassdoor. This data set is collected from Kaggle. The URL to the data set is https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor.

This dataset contains data science job roles and descriptions in different companies and approximate salaries offered for various positions in other companies. This data set contains all the information on data science job posting on Glassdoor, like company name, industry type, sector, minimum and maximum salary offered, location of the company, skills required, and job description of the role. It has all the details of data science job postings in different states of the US. The dataset contains 671 rows and 31 columns, shown below.

Field Name	**Data Description**	**Example Values**
Job Title	Describes the title of the job, Data Type: Char	Data Scientist, Data modeler, Data Analyst, Business Intelligence Analyst, etc
Salary Estimation	Shows the approximate salary range for that job Data Type: Numerical	$137K-171K $75K-131K
Job Description	Includes the full description of that job Data type: Char	Description, Duties & Responsibilities
Rating	Shows the rating of the post out of 5	3.1, 4.2, 3.8, etc

	Data Type: Numerical	
Company Name	Shows the name of the company Data Type: Char	Healthfirst, ManTech, Facebook, etc
Location	Shows where the company is located Data Type: Char	New York, NY, Chantilly, VA, etc
Headquarters	Shows the location of the headquarters of the company Data Type: Char	New York, NY, Herndon, VA, etc
Size	It shows the total number of employees in that company. Data Type: Char	1001 to 5000 employees, 5001 to 10000 employees, etc
Education Level	Describes the education qualification required for this job	Ph.D., Bachelor, Master, etc
Type of Ownership	Describes the company type I,e, nonprofit/public/private/government Data Type: Char	Nonprofit Organizations, Private Practice / Firms, etc.
Industry	Describes the type of industry the applicant will work in Data Type: Char	Research & Development, Insurance, Advertising & Marketing, Enterprise Software & Network Solutions, etc
Sector	Describes the sector of industry the applicant will work in Data Type: Char	Retail, Manufacturing, Government, etc

Revenue	Describes the overall revenue of the company Data Type: Numeric	$100 million (USD), $500 million (USD), etc.
Competition	Describes the competitive company Data Type: Char	EmblemHealth, UnitedHealth Group, Aetna
Minimum Salary	Describes the minimum salary offered for the position Data Type: Numeric	$137000, $75000, $99000, etc
Maximum Salary	Describes the maximum salary offered for the position Data Type: Numeric	$171000, $131000, $132000, etc
Average Salary	Describes the average salary offered for that position Data Type: Numeric	$131000, $171000, $154000, etc
Job State	Describes the state where the applicant will work	NY, VA, CA, etc
Founded	Shows the year the company was founded Data Type: Numeric	1998, 2017, 1985, etc
Python	It describes whether knowledge of Python is required for this job. Data Type: Char	Yes, No
Excel	It describes whether knowledge of	Yes, No

	Excel is required for this job. Data Type: Char	
Hadoop	It describes whether knowledge of Hadoop is required for this job. Data Type: Char	Yes, No
Spark	It describes whether knowledge of Spark is required for this job. Data Type: Char	Yes, No
AWS	It describes whether the knowledge of AWS is required for this job. Data Type: Char	Yes, No
Tableau	It describes whether knowledge of Tableau is required for this job. Data Type: Char	Yes, No
Big Data	It describes whether knowledge of Big Data is required for this job. Data Type: Char	Yes, No
Base Salary	Shows the base salary of the job Data Type: Numerical	$135000, $138000, $160000, etc.
Seniority	Describes the level of the job, such as fresher, senior, etc Data Type: Char	Fresher, senior, etc

Table 1: Data Description

Data Cleaning

Data cleaning is the process of identifying and fixing problems in a dataset. The purpose of data cleansing is to correct data that are inaccurate, incomplete, malformed, duplicated, or irrelevant to the purpose of the dataset. This is typically achieved by replacing, modifying, or deleting data that falls into one of these categories. The data will likely be duplicated or mislabeled when combining multiple data sources. If the data is wrong, the results and algorithms are unreliable, even if they look correct. Since the process is different for each dataset, there is no absolute way to indicate the exact steps of the data cleansing process. Our decisions are usually based on datasets, so if the data quality is poor, our results will not be accurate. Therefore, data cleaning is essential because you can get high-quality data that leads to better-quality decisions. Not all data is good data in the data set. There was a little junk data. The dataset used for this analysis contained some null values. Some datasets had blank / missing values, so these data were deleted and filtered while focusing on the required data set. Unnecessary columns were removed, and a few cues were split. Below are the few steps taken to clean the dataset.

Data Cleaning Steps

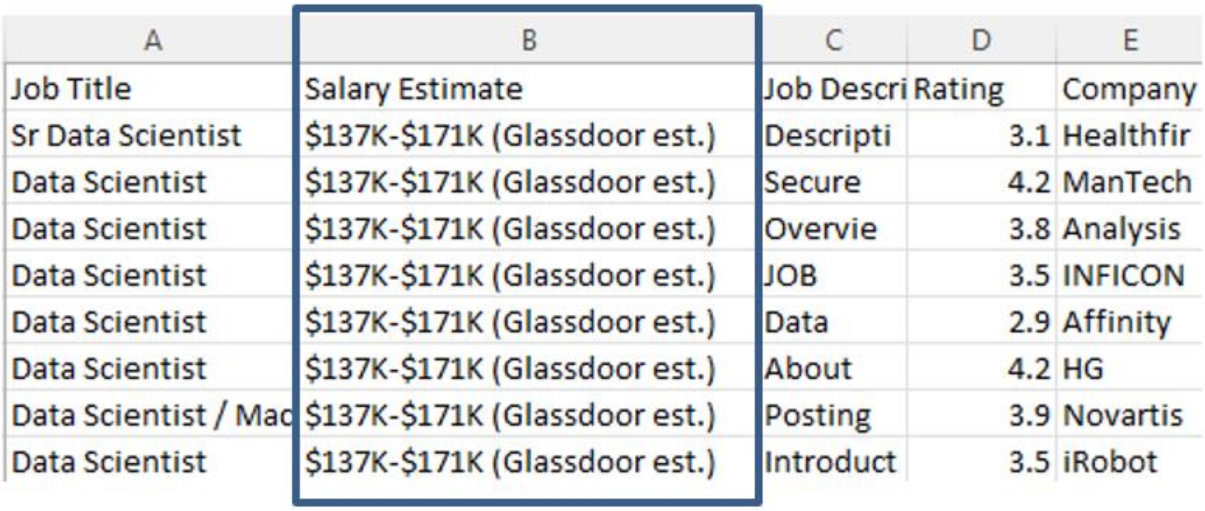
1. **Data cleaning category name:** Extract the required information from the column Salary_Estimation and remove the unwanted information

Steps to clean the Data:

- (Glassdoor. est) was removed from the column
- Formula Right(text, num_char) and LEFT(text, num_char) was used to extract the minimum and maximum value from the field
- Remove ‘$’ and ‘K’ from the number and add Zeros at the end.

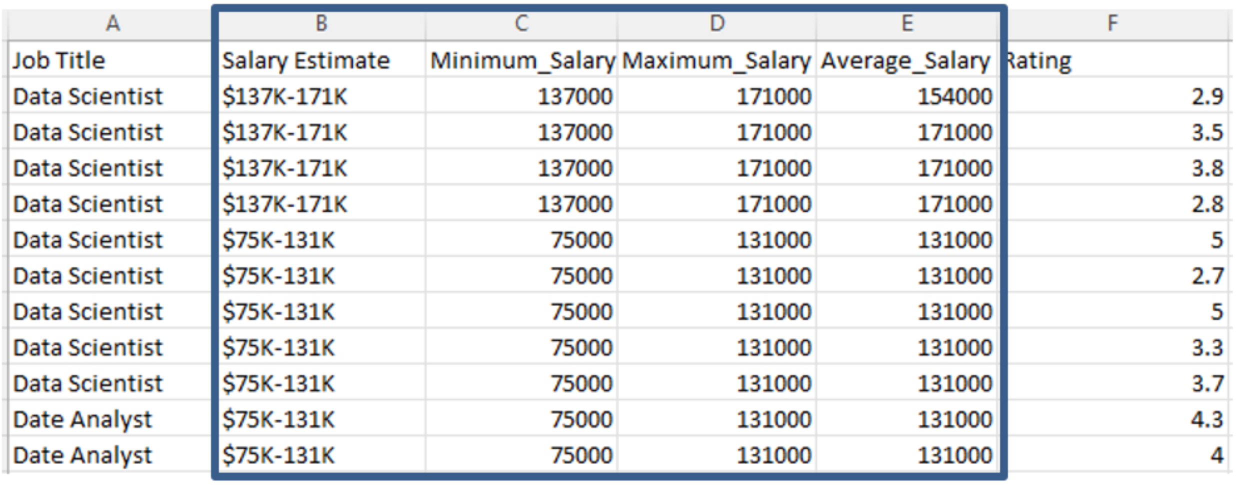
- Added minimum and maximum salary to a new column
- Average formula was used to get the average salary

Sample data set before cleaning



A	B	C	D	E
Job Title	Salary Estimate	Job Descri	Rating	Company
Sr Data Scientist	$137K-$171K (Glassdoor est.)	Descripti	3.1	Healthfir
Data Scientist	$137K-$171K (Glassdoor est.)	Secure	4.2	ManTech
Data Scientist	$137K-$171K (Glassdoor est.)	Overvie	3.8	Analysis
Data Scientist	$137K-$171K (Glassdoor est.)	JOB	3.5	INFICON
Data Scientist	$137K-$171K (Glassdoor est.)	Data	2.9	Affinity
Data Scientist	$137K-$171K (Glassdoor est.)	About	4.2	HG
Data Scientist / Mac	$137K-$171K (Glassdoor est.)	Posting	3.9	Novartis
Data Scientist	$137K-$171K (Glassdoor est.)	Introduct	3.5	iRobot

Sample data set after cleaning



A	B	C	D	E	F
Job Title	Salary Estimate	Minimum_Salary	Maximum_Salary	Average_Salary	Rating
Data Scientist	$137K-171K	137000	171000	154000	2.9
Data Scientist	$137K-171K	137000	171000	171000	3.5
Data Scientist	$137K-171K	137000	171000	171000	3.8
Data Scientist	$137K-171K	137000	171000	171000	2.8
Data Scientist	$75K-131K	75000	131000	131000	5
Data Scientist	$75K-131K	75000	131000	131000	2.7
Data Scientist	$75K-131K	75000	131000	131000	5
Data Scientist	$75K-131K	75000	131000	131000	3.3
Data Scientist	$75K-131K	75000	131000	131000	3.7
Date Analyst	$75K-131K	75000	131000	131000	4.3
Date Analyst	$75K-131K	75000	131000	131000	4

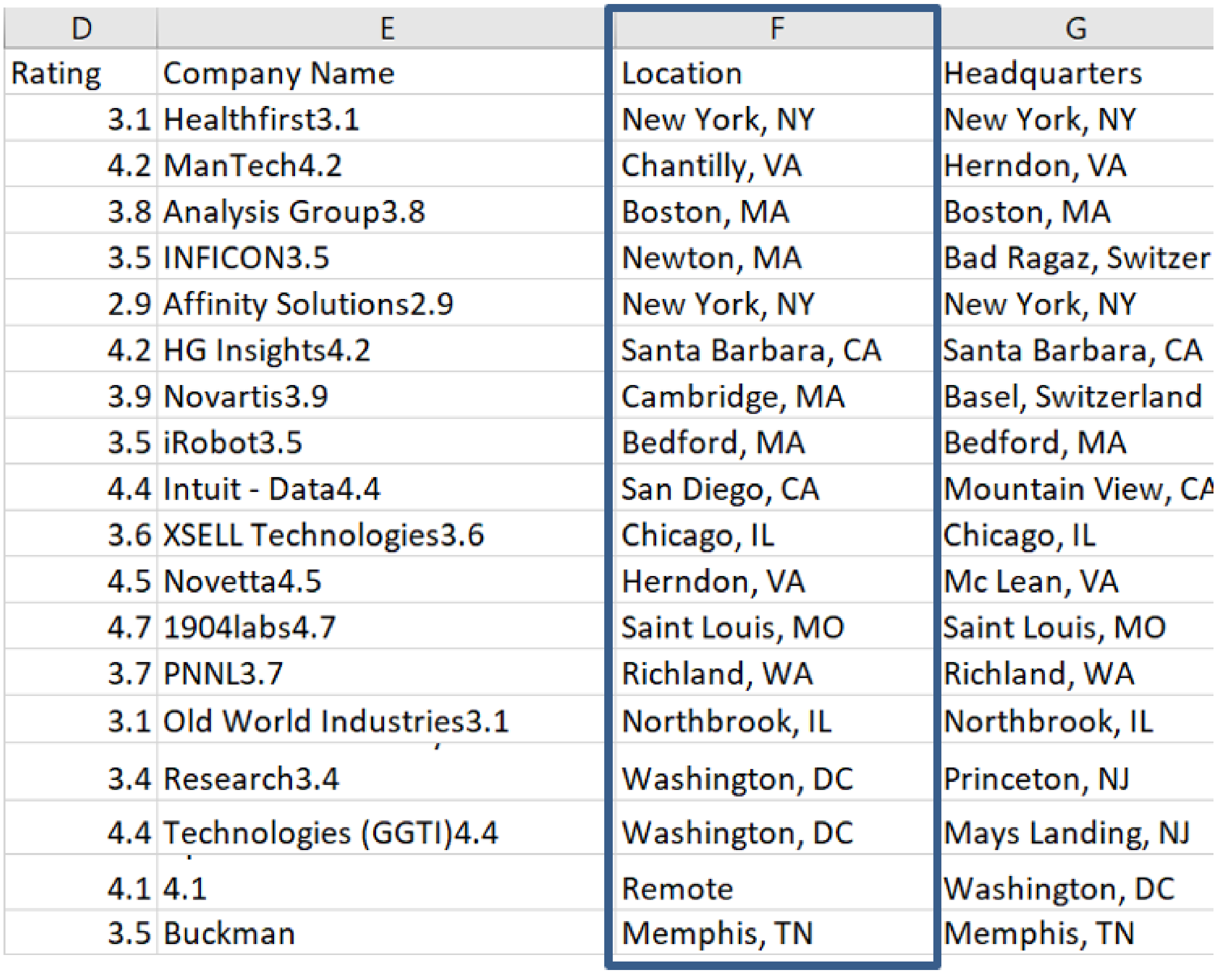
2. **Data cleaning category name:** Splitting the column

Steps to clean the Data

The column location had city and state information. This information was split into two-column I,e Location and Job_State

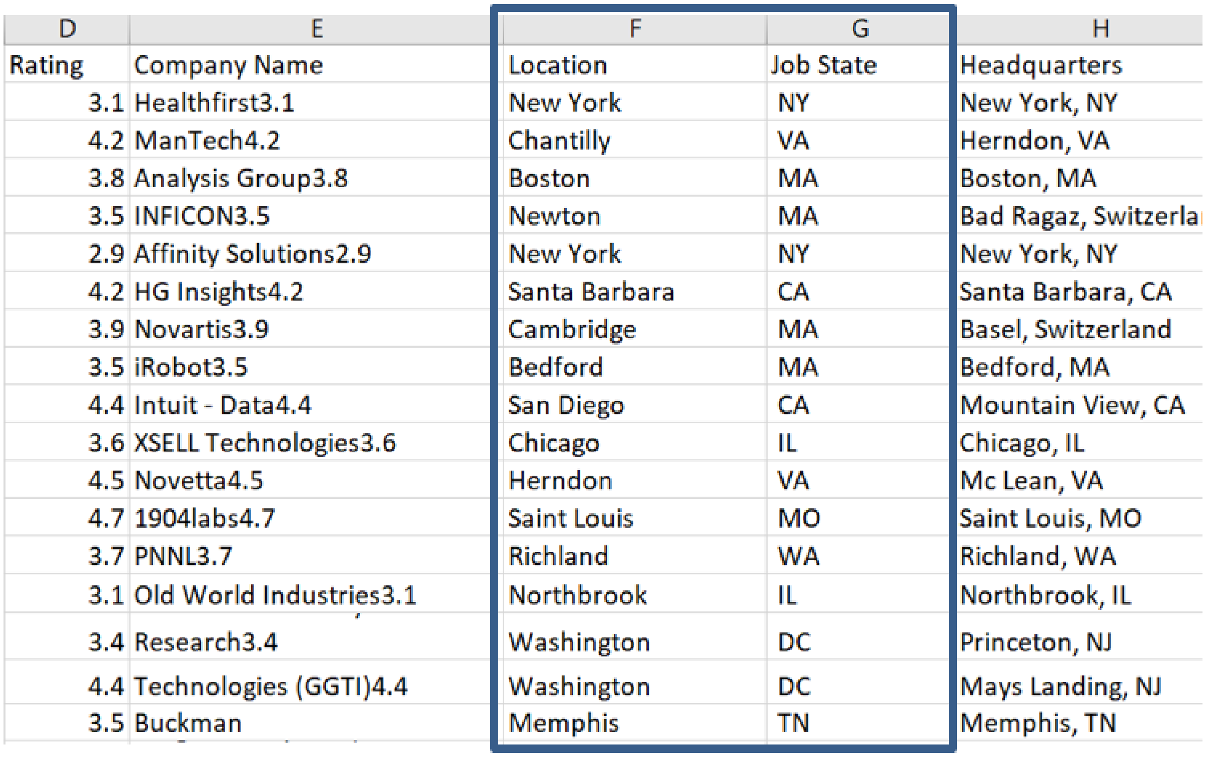
- Text to the column was used to split the job city and state

Sample data set before cleaning



D	E	F	G
Rating	Company Name	Location	Headquarters
3.1	Healthfirst3.1	New York, NY	New York, NY
4.2	ManTech4.2	Chantilly, VA	Herndon, VA
3.8	Analysis Group3.8	Boston, MA	Boston, MA
3.5	INFICON3.5	Newton, MA	Bad Ragaz, Switzer
2.9	Affinity Solutions2.9	New York, NY	New York, NY
4.2	HG Insights4.2	Santa Barbara, CA	Santa Barbara, CA
3.9	Novartis3.9	Cambridge, MA	Basel, Switzerland
3.5	iRobot3.5	Bedford, MA	Bedford, MA
4.4	Intuit - Data4.4	San Diego, CA	Mountain View, CA
3.6	XSELL Technologies3.6	Chicago, IL	Chicago, IL
4.5	Novetta4.5	Herndon, VA	Mc Lean, VA
4.7	1904labs4.7	Saint Louis, MO	Saint Louis, MO
3.7	PNNL3.7	Richland, WA	Richland, WA
3.1	Old World Industries3.1	Northbrook, IL	Northbrook, IL
3.4	Research3.4	Washington, DC	Princeton, NJ
4.4	Technologies (GGTI)4.4	Washington, DC	Mays Landing, NJ
4.1	4.1	Remote	Washington, DC
3.5	Buckman	Memphis, TN	Memphis, TN

Sample data set after cleaning



D	E	F	G	H
Rating	Company Name	Location	Job State	Headquarters
3.1	Healthfirst3.1	New York	NY	New York, NY
4.2	ManTech4.2	Chantilly	VA	Herndon, VA
3.8	Analysis Group3.8	Boston	MA	Boston, MA
3.5	INFICON3.5	Newton	MA	Bad Ragaz, Switzerla
2.9	Affinity Solutions2.9	New York	NY	New York, NY
4.2	HG Insights4.2	Santa Barbara	CA	Santa Barbara, CA
3.9	Novartis3.9	Cambridge	MA	Basel, Switzerland
3.5	iRobot3.5	Bedford	MA	Bedford, MA
4.4	Intuit - Data4.4	San Diego	CA	Mountain View, CA
3.6	XSELL Technologies3.6	Chicago	IL	Chicago, IL
4.5	Novetta4.5	Herndon	VA	Mc Lean, VA
4.7	1904labs4.7	Saint Louis	MO	Saint Louis, MO
3.7	PNNL3.7	Richland	WA	Richland, WA
3.1	Old World Industries3.1	Northbrook	IL	Northbrook, IL
3.4	Research3.4	Washington	DC	Princeton, NJ
4.4	Technologies (GGTI)4.4	Washington	DC	Mays Landing, NJ
3.5	Buckman	Memphis	TN	Memphis, TN

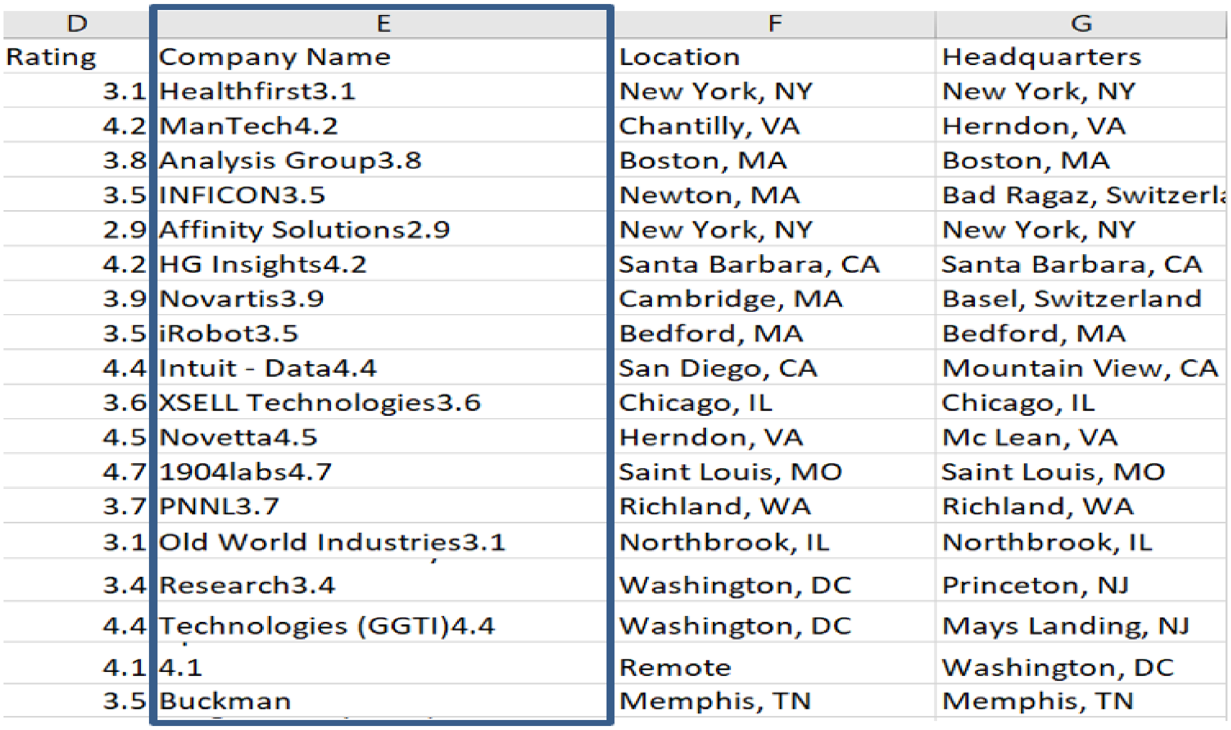
3. **Data cleaning category name:** Removing irrelevant observation

Steps to clean the Data

The column company_name has a few junk numbers at the end of the company name. These numbers were removed from the column.

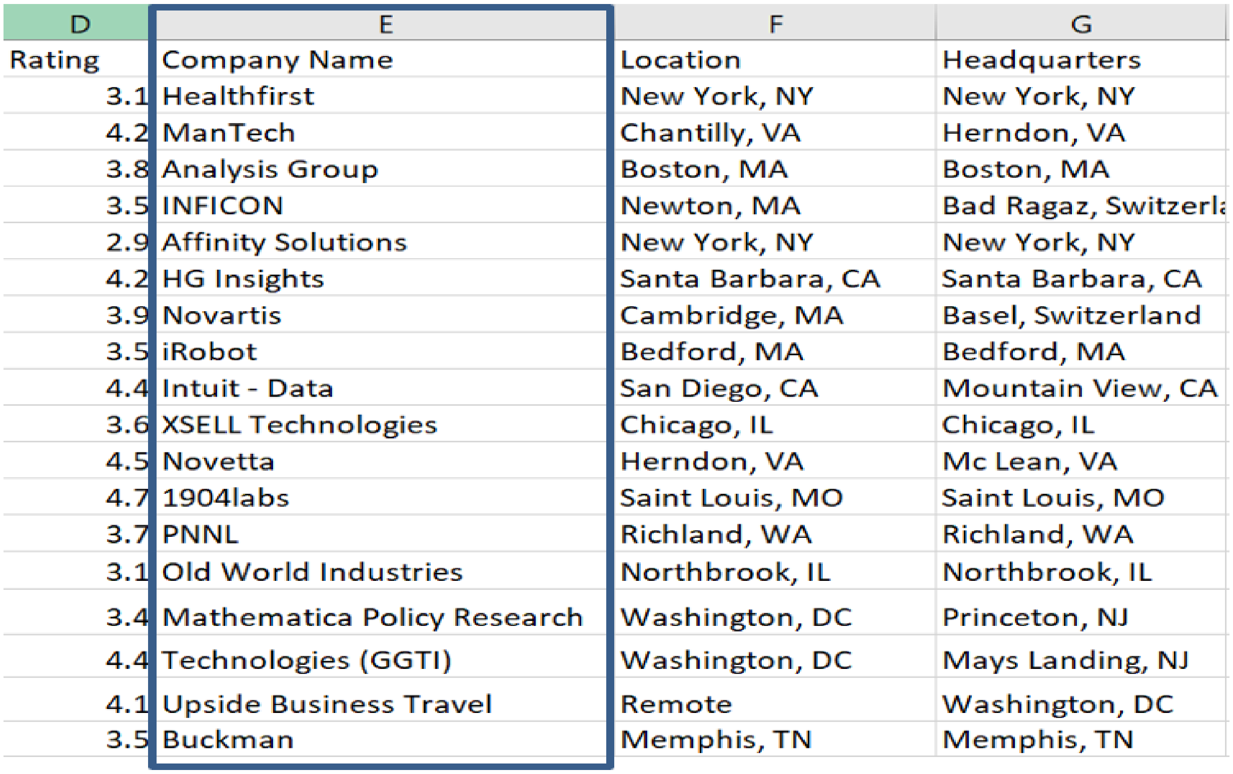
- Using the formula REPLACE(text,LEN(text),num_chars,new_text) to remove the last three characters from the field.

Sample data set before cleaning



D	E	F	G
Rating	Company Name	Location	Headquarters
3.1	Healthfirst3.1	New York, NY	New York, NY
4.2	ManTech4.2	Chantilly, VA	Herndon, VA
3.8	Analysis Group3.8	Boston, MA	Boston, MA
3.5	INFICON3.5	Newton, MA	Bad Ragaz, Switzerl
2.9	Affinity Solutions2.9	New York, NY	New York, NY
4.2	HG Insights4.2	Santa Barbara, CA	Santa Barbara, CA
3.9	Novartis3.9	Cambridge, MA	Basel, Switzerland
3.5	iRobot3.5	Bedford, MA	Bedford, MA
4.4	Intuit - Data4.4	San Diego, CA	Mountain View, CA
3.6	XSELL Technologies3.6	Chicago, IL	Chicago, IL
4.5	Novetta4.5	Herndon, VA	Mc Lean, VA
4.7	1904labs4.7	Saint Louis, MO	Saint Louis, MO
3.7	PNNL3.7	Richland, WA	Richland, WA
3.1	Old World Industries3.1	Northbrook, IL	Northbrook, IL
3.4	Research3.4	Washington, DC	Princeton, NJ
4.4	Technologies (GGTI)4.4	Washington, DC	Mays Landing, NJ
4.1	4.1	Remote	Washington, DC
3.5	Buckman	Memphis, TN	Memphis, TN

Sample data set after cleaning



D	E	F	G
Rating	Company Name	Location	Headquarters
3.1	Healthfirst	New York, NY	New York, NY
4.2	ManTech	Chantilly, VA	Herndon, VA
3.8	Analysis Group	Boston, MA	Boston, MA
3.5	INFICON	Newton, MA	Bad Ragaz, Switzerl
2.9	Affinity Solutions	New York, NY	New York, NY
4.2	HG Insights	Santa Barbara, CA	Santa Barbara, CA
3.9	Novartis	Cambridge, MA	Basel, Switzerland
3.5	iRobot	Bedford, MA	Bedford, MA
4.4	Intuit - Data	San Diego, CA	Mountain View, CA
3.6	XSELL Technologies	Chicago, IL	Chicago, IL
4.5	Novetta	Herndon, VA	Mc Lean, VA
4.7	1904labs	Saint Louis, MO	Saint Louis, MO
3.7	PNNL	Richland, WA	Richland, WA
3.1	Old World Industries	Northbrook, IL	Northbrook, IL
3.4	Mathematica Policy Research	Washington, DC	Princeton, NJ
4.4	Technologies (GGTI)	Washington, DC	Mays Landing, NJ
4.1	Upside Business Travel	Remote	Washington, DC
3.5	Buckman	Memphis, TN	Memphis, TN

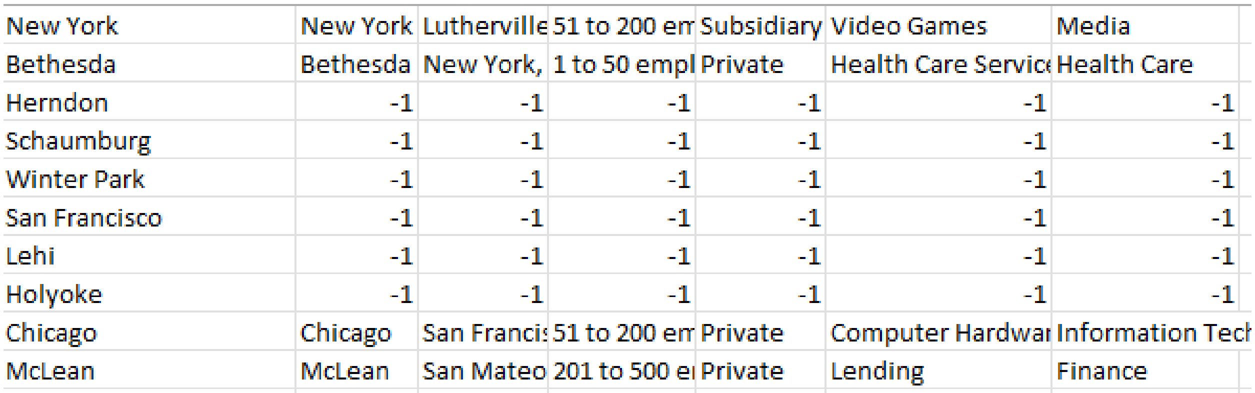
4. **Data cleaning category name:** Filling the null values

Steps to clean the data

A few columns and rows had missing information. This information was manually added to the dataset. For example, a few companies were missing location, headquarters, size, type of ownership, industry, and sector. This information was added to the dataset.

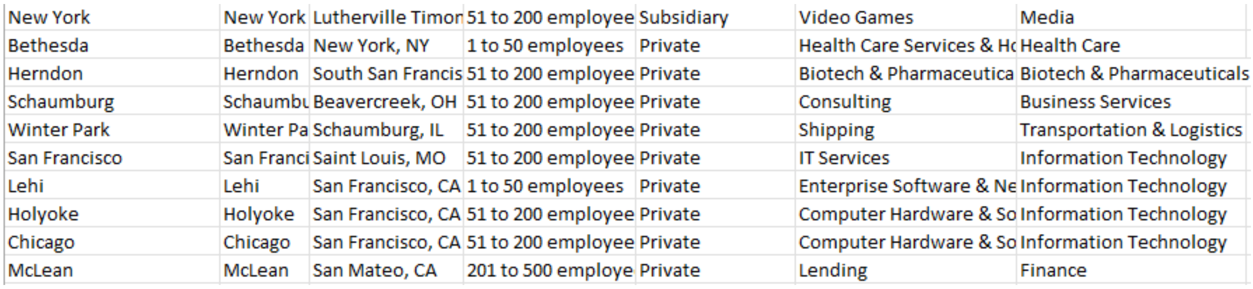
- Added filter to get the blank/ missing values
- Manually missing data was entered

Sample data set before cleaning



New York	New York	Lutherville	51 to 200 em	Subsidiary	Video Games	Media
Bethesda	Bethesda	New York,	1 to 50 empl	Private	Health Care Servic	Health Care
Herndon	-1	-1	-1	-1	-1	-1
Schaumburg	-1	-1	-1	-1	-1	-1
Winter Park	-1	-1	-1	-1	-1	-1
San Francisco	-1	-1	-1	-1	-1	-1
Lehi	-1	-1	-1	-1	-1	-1
Holyoke	-1	-1	-1	-1	-1	-1
Chicago	Chicago	San Francis	51 to 200 em	Private	Computer Hardwar	Information Tech
McLean	McLean	San Mateo	201 to 500 e	Private	Lending	Finance

Sample data set after cleaning



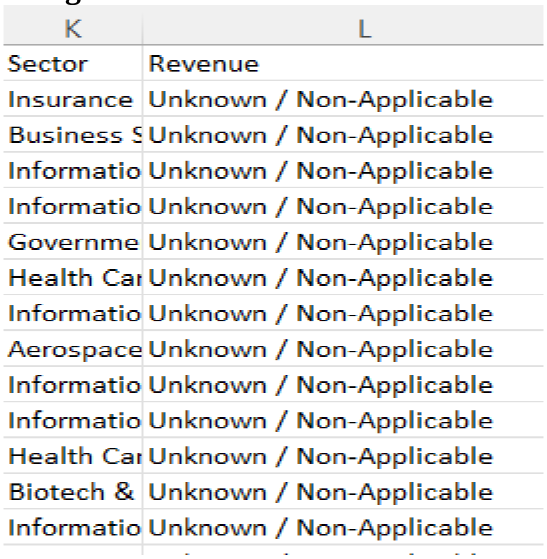
New York	New York	Lutherville Timon	51 to 200 employee	Subsidiary	Video Games	Media
Bethesda	Bethesda	New York, NY	1 to 50 employees	Private	Health Care Services & Hc	Health Care
Herndon	Herndon	South San Francis	51 to 200 employee	Private	Biotech & Pharmaceutica	Biotech & Pharmaceuticals
Schaumburg	Schaumbu	Beavercreek, OH	51 to 200 employee	Private	Consulting	Business Services
Winter Park	Winter Pa	Schaumburg, IL	51 to 200 employee	Private	Shipping	Transportation & Logistics
San Francisco	San Franci	Saint Louis, MO	51 to 200 employee	Private	IT Services	Information Technology
Lehi	Lehi	San Francisco, CA	1 to 50 employees	Private	Enterprise Software & Ne	Information Technology
Holyoke	Holyoke	San Francisco, CA	51 to 200 employee	Private	Computer Hardware & So	Information Technology
Chicago	Chicago	San Francisco, CA	51 to 200 employee	Private	Computer Hardware & So	Information Technology
McLean	McLean	San Mateo, CA	201 to 500 employe	Private	Lending	Finance

5. **Data cleaning category name:** Removing Unknows values

Steps for cleaning the data

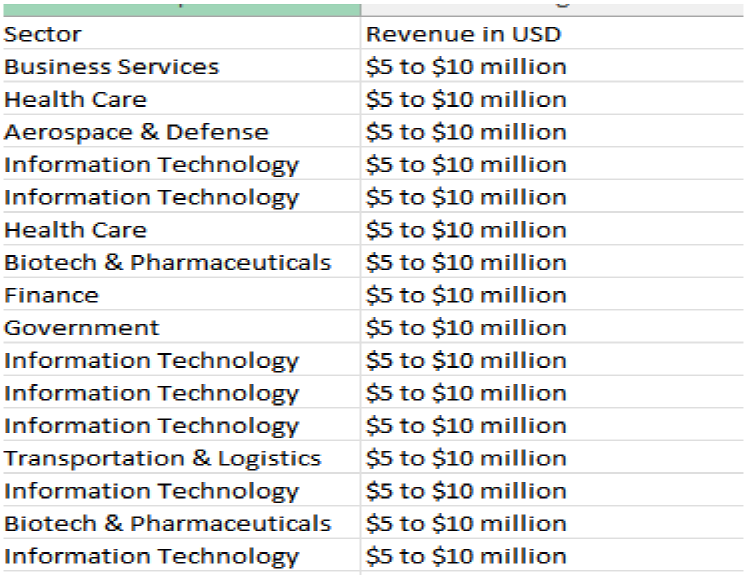
- Added filter to select the unknown / Non-Applicable. These fields were deleted and kept null
- Many other irrelevant data were deleted. Index row, description, competitors, same_state, company_age, and job_simp were deleted from the data set to make it a more readable dataset.

Sample data set before cleaning



K	L
Sector	Revenue
Insurance	Unknown / Non-Applicable
Business S	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable
Governme	Unknown / Non-Applicable
Health Ca	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable
Aerospace	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable
Health Ca	Unknown / Non-Applicable
Biotech &	Unknown / Non-Applicable
Informatio	Unknown / Non-Applicable

Sample data set after cleaning

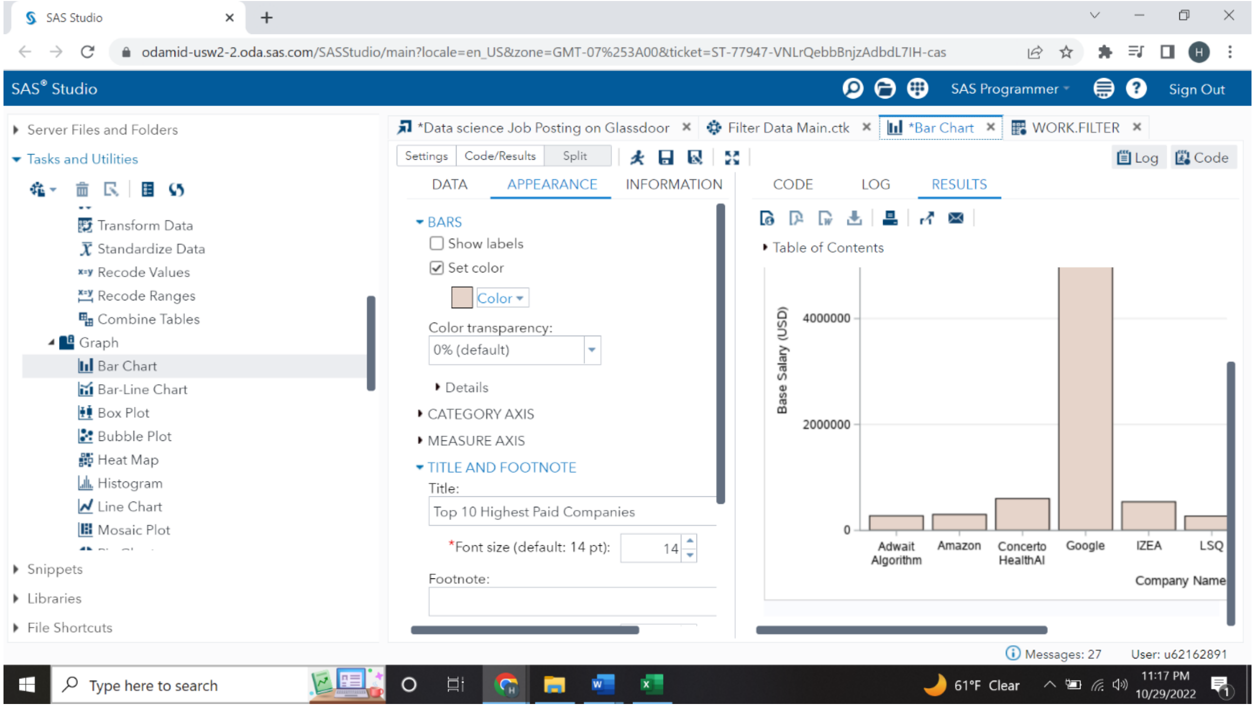


Sector	Revenue in USD
Business Services	$5 to $10 million
Health Care	$5 to $10 million
Aerospace & Defense	$5 to $10 million
Information Technology	$5 to $10 million
Information Technology	$5 to $10 million
Health Care	$5 to $10 million
Biotech & Pharmaceuticals	$5 to $10 million
Finance	$5 to $10 million
Government	$5 to $10 million
Information Technology	$5 to $10 million
Information Technology	$5 to $10 million
Information Technology	$5 to $10 million
Transportation & Logistics	$5 to $10 million
Information Technology	$5 to $10 million
Biotech & Pharmaceuticals	$5 to $10 million
Information Technology	$5 to $10 million

Data Analysis and Visualization

Data visualization is a graphical representation of information and data. Using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in your data. Data visualization tools and technologies areessential to analyze vast amounts of information and making informed decisions in big data.

1. The top ten highest-paid companies for data scientist job roles



odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-07%253A00&ticket=ST-77947-VNLrQebbBnjzAdbdL7IH-cas

BARS

☐ Show labels

☑ Set color

Color transparency: 0% (default)

TITLE AND FOOTNOTE

Title: Top 10 Highest Paid Companies

*Font size (default: 14 pt): 14

Footnote:

Messages: 27 User: u62162891

Figure 1: Screenshot of the output

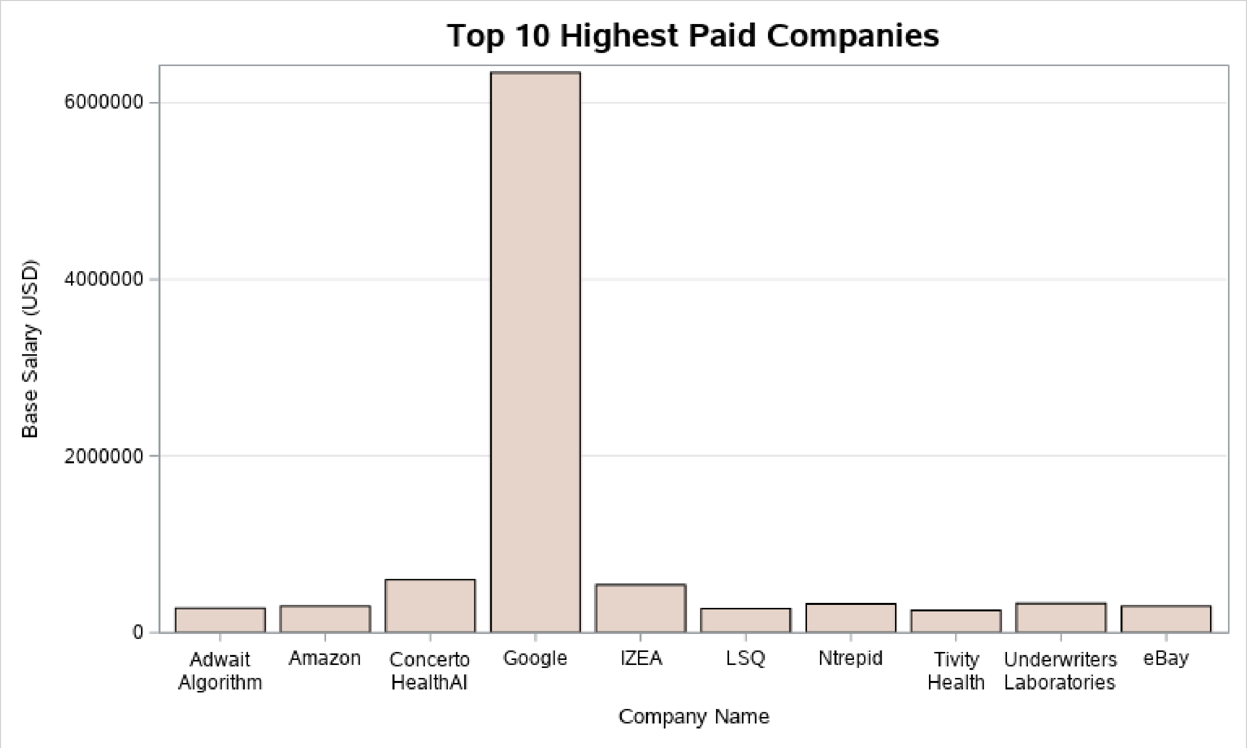
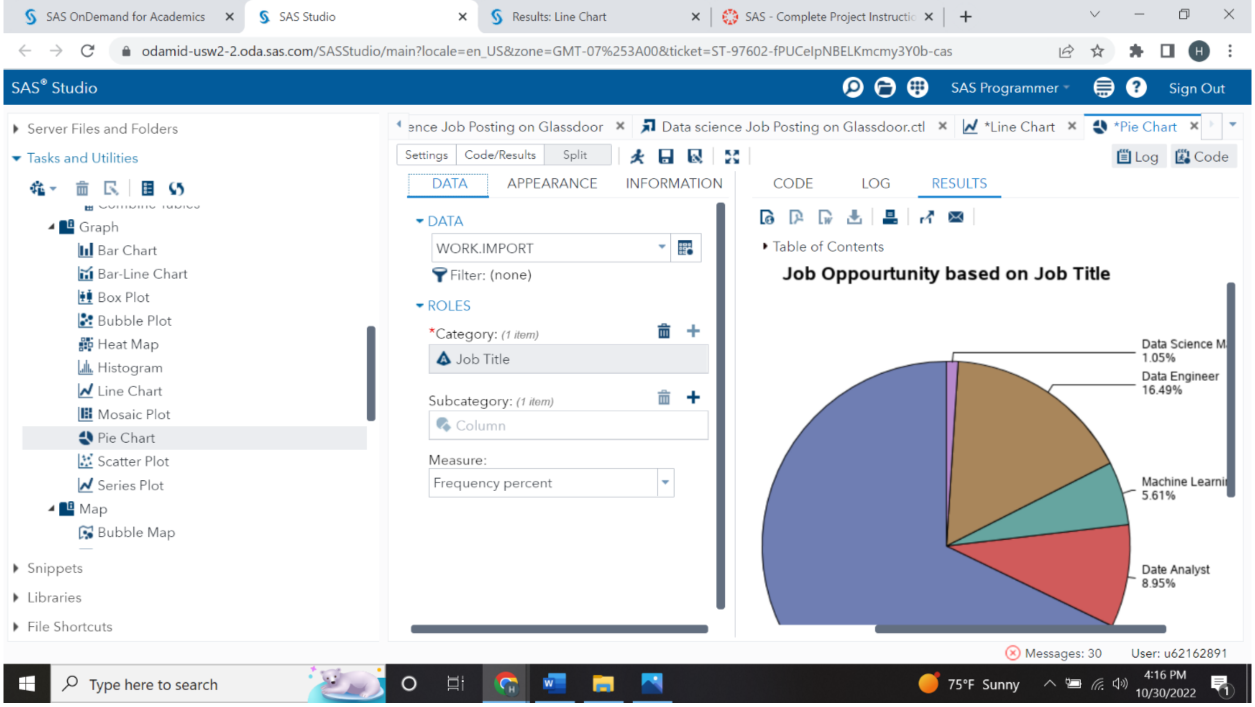


Figure 2: Bar graph that shows the top 10 highest-paid companies

This analysis helps us understand the salaries offered by different companies. Different companies offer different pay ranges for various job roles. Job salary is one of the main criteria when applying for a job. With this analysis, we can find the salary offered by different companies.

The above bar graph shows the top ten companies with the highest salaries. The graph's x-axis is the company name, and the chart's y-axis is the Base salary in USD. The filter is added to the salary column to get the top ten highest salaries. The above graph shows that Google has the highest salary offered, followed by Concerto HealthAl, IZEA, etc. It is seen that the salary provided by Google is much higher than the average salary of all ten companies. There is a significant difference in wages provided by the top tenth and the rest of the companies. If we compare the result of the other nine companies, there is not much difference in the range of salary offered. So, from the analysis, we can predict that Google employees receive higher pay than the other companies.

2. The job opportunities for different job roles in the USA



DATA
WORK.IMPORT
Filter: (none)

ROLES
*Category: (1 item)
Job Title

Subcategory: (1 item)
Column

Measure:
Frequency percent

Figure 3: Screenshot of the output

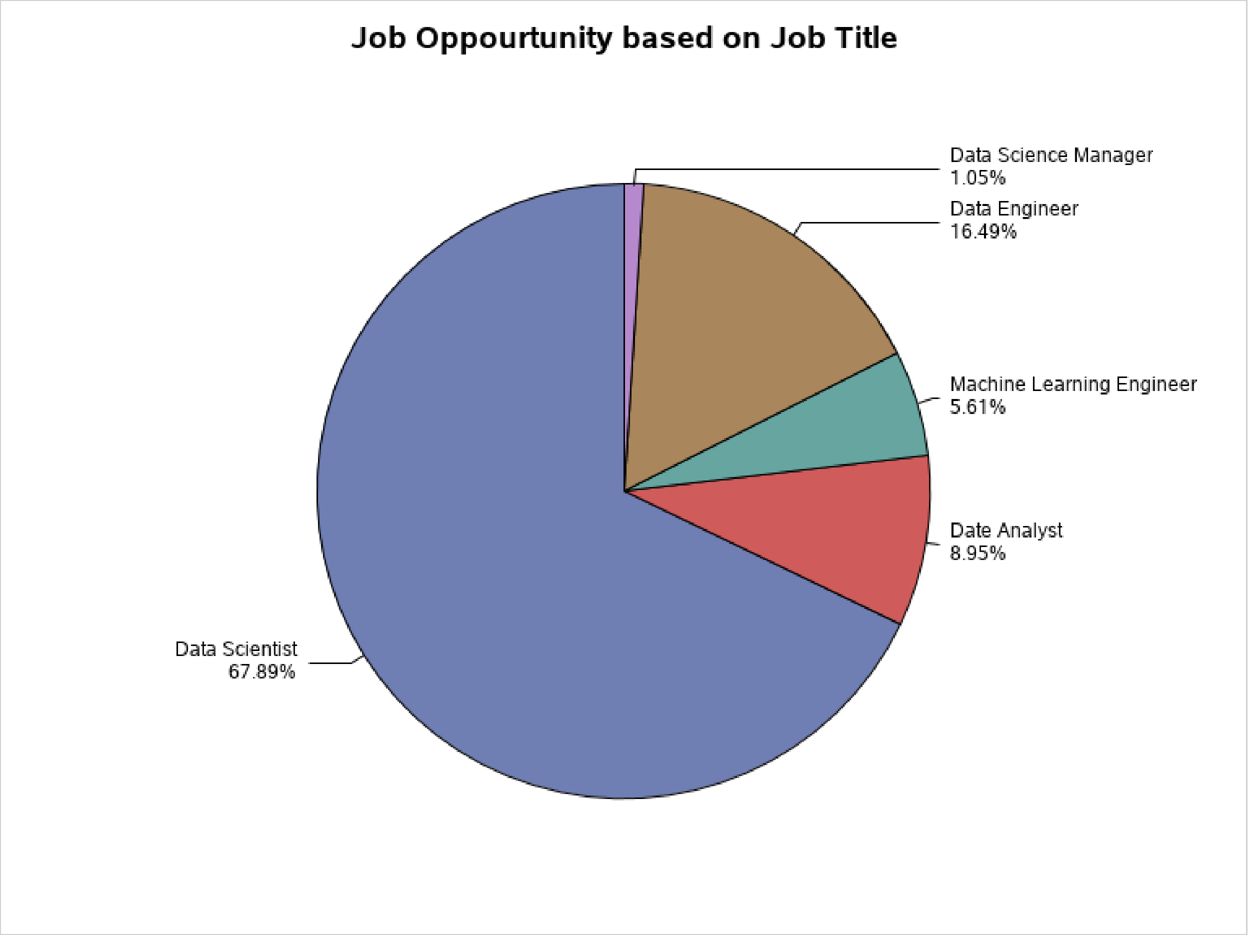
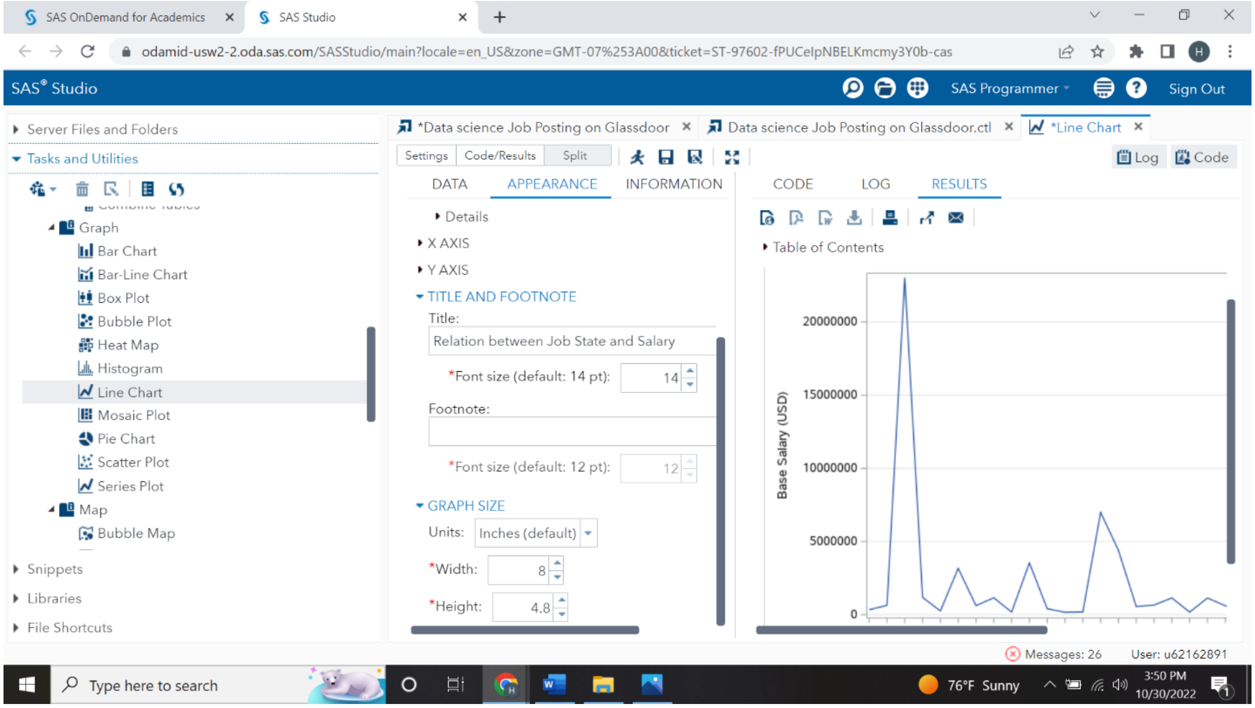


Figure 4: Pie chart showing the job opportunities for different job roles in the USA

The above analysis represents the top five job roles with the highest opportunities. A pie chart is used to describe this analysis. From the result, we can observe that Data Scientist has the highest number of job opportunity, followed by Data engineering and Data Analyst. The top five job roles with the highest number of options are Data Scientist, Data Engineering, Data Analyst, Machine Learning engineering, and Data Scientist Manager.

From the analysis, we can conclude that the candidate applying for Data Scientist job roles may find many job openings in the USA. This analysis will help us understand the job opportunity for different parts of the USA. If a candidate is interested in only one job role, then the Job titles can be filtered, and only the roles we are interested in can be analyzed. With the help of this chart, we can see which job role has the highest job opportunity; based on that; a candidate can explore which job role they want to apply for.

3. **Analyze the salary offered based on Location. Does Job location play any role in the salaries?**



Title: Relation between Job State and Salary

*Font size (default: 14 pt): 14

Footnote:

*Font size (default: 12 pt): 12

Units: Inches (default)

*Width: 8

*Height: 4.8

Figure 5: Screenshot of the output

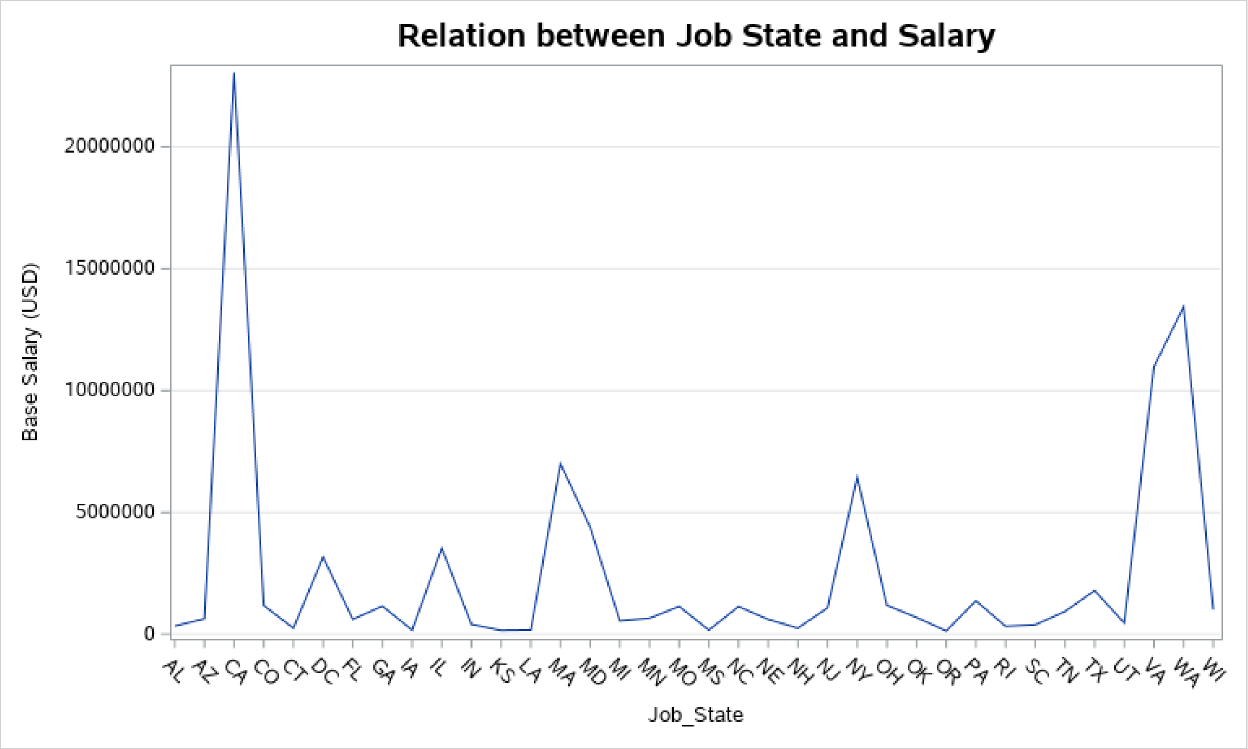


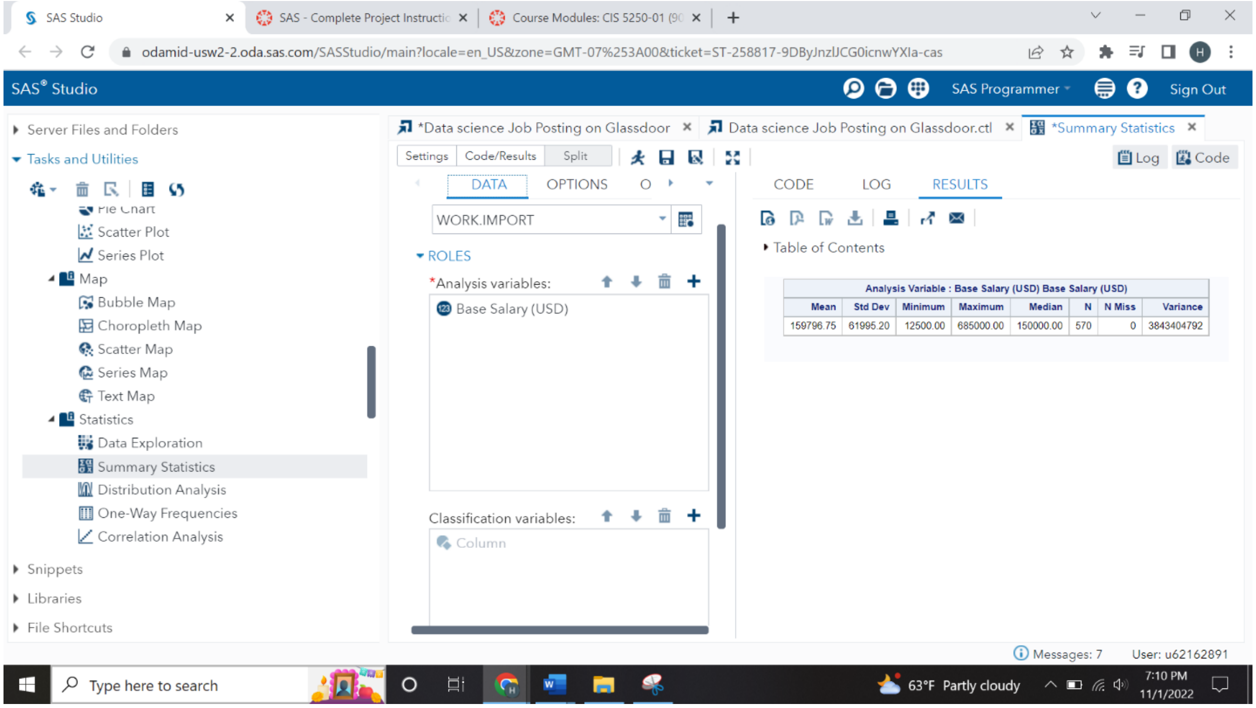
Figure 6: Line chart that shows the relationship between the job state and base salary

The above line chart represents the salary offered by different states in the USA. Location is another criteria to consider while applying for a job. Few states will have higher job opportunities with higher pay, and few states will have lower job opportunities with lower pay. With this analysis, we can see the salary offered based on the location.

The X-axis represents the job state, and the y-axis represents the base salary in UDS. As per the graph, it is seen that CA, that is, California state has the highest pay, and the second state with the highest pay is Washington. From the analysis, we can see that different locations have different salaries, and location is one of the criteria to consider when applying for jobs.

Statistical Summary

1. Statistical analysis for the variable Base Salary (USD)



WORK.IMPORT

ROLES

*Analysis variables:

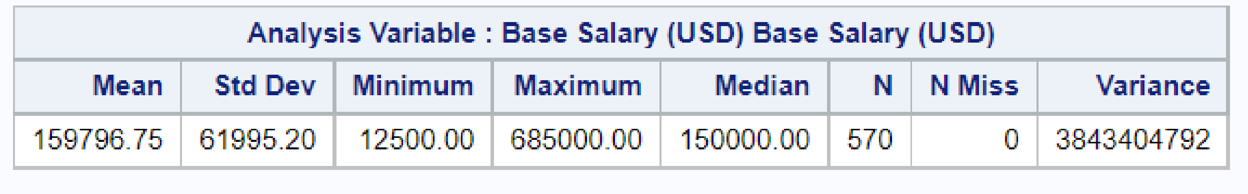
Base Salary (USD)

Classification variables:

Column

Analysis Variable : Base Salary (USD) Base Salary (USD)							
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Variance
159796.75	61995.20	12500.00	685000.00	150000.00	570	0	3843404792

Figure 7: Screenshot of the output



Analysis Variable : Base Salary (USD) Base Salary (USD)							
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Variance
159796.75	61995.20	12500.00	685000.00	150000.00	570	0	3843404792

Figure 8: Summary statistics for the variable base salary (USD)

Mean:

The mean of the base salary is 159796.75 UDS. The minimum salary is 12500, and the maximum salary is 685000 USD. The data indicated that a candidate's average salary is around 159796.75 UDS. This mean value indicated an average salary offered to a candidate compared to different companies. This mean value can be compared against an individual base salary to judge if they have comparably provided a good salary compared to another candidate. Similarly, their respective means can be used to compare the base salary of a smaller group against a more comprehensive group. An example would be comparing the mean base salary offered by a small company and a big company.

Standard Deviation:

The standard deviation of the variable base salary is 61995. 20. That means that each company's base salary is at an average difference of 61995.20 from the mean salary of all the companies. This value shows how the data is spread out from the mean. The standard deviation is less than the mean value, indicating that the data points were below the mean. Since the standard deviation is lower than the mean value, the data are more clustered around the mean.

Median

The median of the variable base salary is 150000, and the mean value is 159796.75. The median provides a helpful measure of the center of a dataset. We can see that the median value is very close to the mean value. Since the median value is almost very close to the mean, we can conclude that the dataset distribution is symmetrical. We are comparing the median to the mean; we can say that the data set is evenly distributed from the lowest to highest values.

Minimum

We see the minimum base salary offered to a job. When we compare the base salary provided by different companies for different job roles and different years of experience, we see that the minimum salary offered is $12500

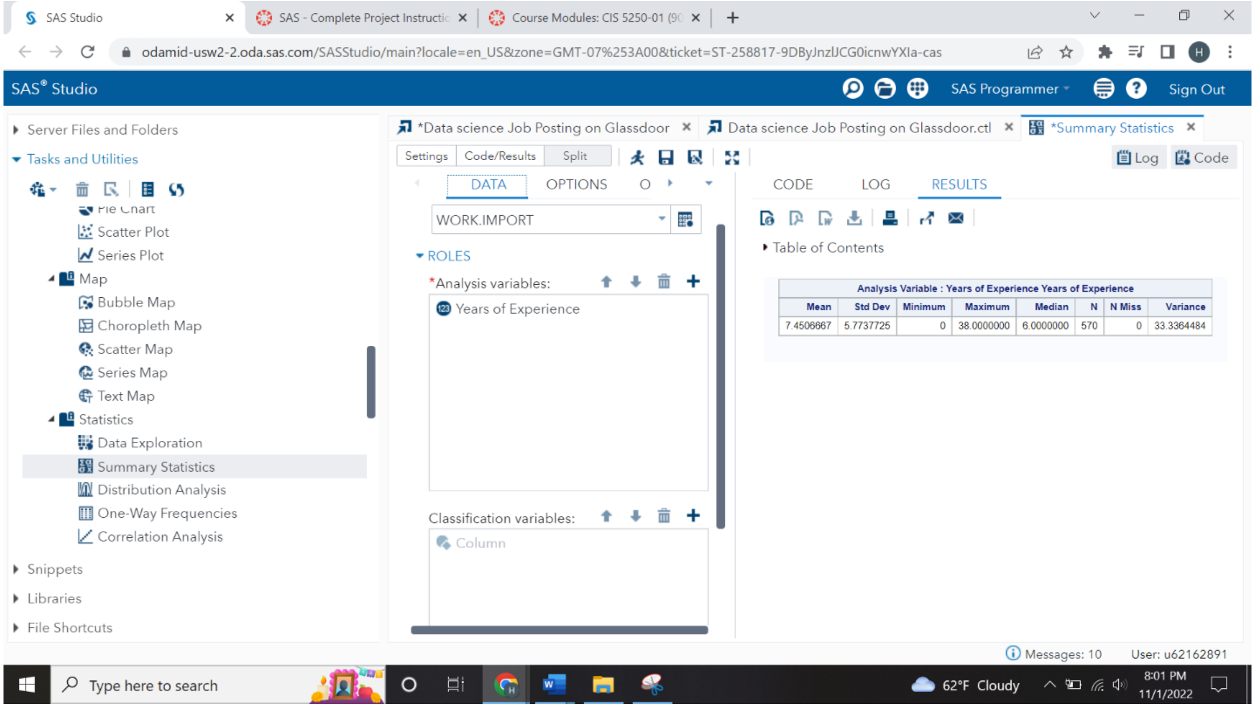
Maximum

We see the maximum base salary offered to a job. When we compare the base salary provided by different companies for different job roles and different years of experience, we see that the maximum salary offered is $685000

Variance

The variance of the variable base salary is 3843404792, which is very high. This indicates that some companies offer very high salaries, and some offer very low salaries. This also suggests that salary values are more spread out.

2. Statical analysis for the variable years of experience

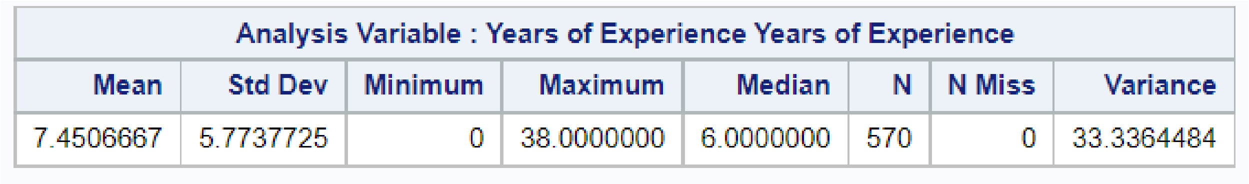


WORK.IMPORT

Analysis variables: Years of Experience

Analysis Variable : Years of Experience Years of Experience							
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Variance
7.4506667	5.7737725	0	38.0000000	6.0000000	570	0	33.3364484

Figure 9: Screenshot of the output



Analysis Variable : Years of Experience Years of Experience							
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Variance
7.4506667	5.7737725	0	38.0000000	6.0000000	570	0	33.3364484

Figure 10: Summary statistics for the variable years of experience (USD)

Mean:

The mean of the variable year of experience is 7.45 UDS. The minimum experience required is 0 years, and the maximum experience is 38 years. This data indicated that the job range starts from the freshers' level and goes to the seniority level. The mean value indicates the average years of experience most companies require for the job. Since the data set has a wide range from fresher to seniority levels, the mean value will help us compare how many years of experience are required to get into different job positions.

Standard Deviation:

The standard deviation of the variable year of experience is 5.77. The standard deviation is less than the mean value, which indicates that the data points were below the mean. Since the standard deviation is lower than the mean value, the data are more clustered around the mean. This value shows how the data is spread out from the mean.

Median

The median of the variable year of experience is 6, and the mean value is 7.45. We can see that the median value is close to the mean value. We are comparing the median to the mean; we can say that the data set is evenly distributed from the lowest to highest values.

Minimum

The minimum years of experience required for the job are zero. This data set has different job level requirements starting from freshers and increasing to seniority level. We see that the minimum year of experience required for the job for different companies and for different job roles is zero. This data set has the job requirements for freshers with zero experience.

Maximum

The maximum number of years of experience required for the job is 38 years. We see the maximum year of experience needed for the job for different companies and for various job roles is 38 years.

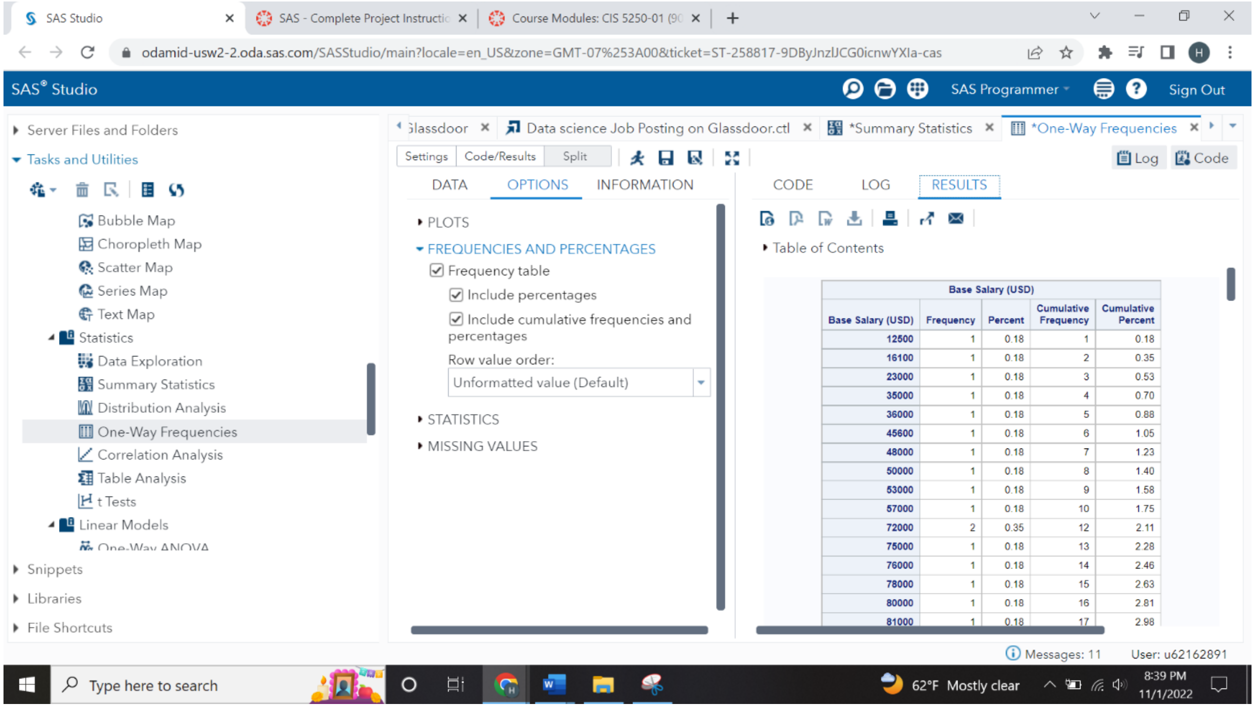
Variance

The variance of the variable year of experience is 33.33, which is less than the maximum value but higher than the mean. This indicates that some companies require higher experience candidates, and some are looking for freshers. This also suggests that years of experience in job roles are more spread out.

Statistical Test

1. One-way frequency

One-way frequency refers to a tabulation of the data which only examines one categorical variable at a time. The frequency can tabulate this simple structure and produce tests for equal proportions across the categories. Below is the one-way frequency result for the variable Base Salary (USD)



☑ Frequency table

☑ Include percentages

☑ Include cumulative frequencies and percentages

Row value order:
Unformatted value (Default)

Base Salary (USD)				
Base Salary (USD)	Frequency	Percent	Cumulative Frequency	Cumulative Percent
12500	1	0.18	1	0.18
16100	1	0.18	2	0.35
23000	1	0.18	3	0.53
35000	1	0.18	4	0.70
36000	1	0.18	5	0.88
45600	1	0.18	6	1.05
48000	1	0.18	7	1.23
50000	1	0.18	8	1.40
53000	1	0.18	9	1.58
57000	1	0.18	10	1.75
72000	2	0.35	12	2.11
75000	1	0.18	13	2.28
76000	1	0.18	14	2.46
78000	1	0.18	15	2.63
80000	1	0.18	16	2.81
81000	1	0.18	17	2.98

Figure 11: Screenshot of the output

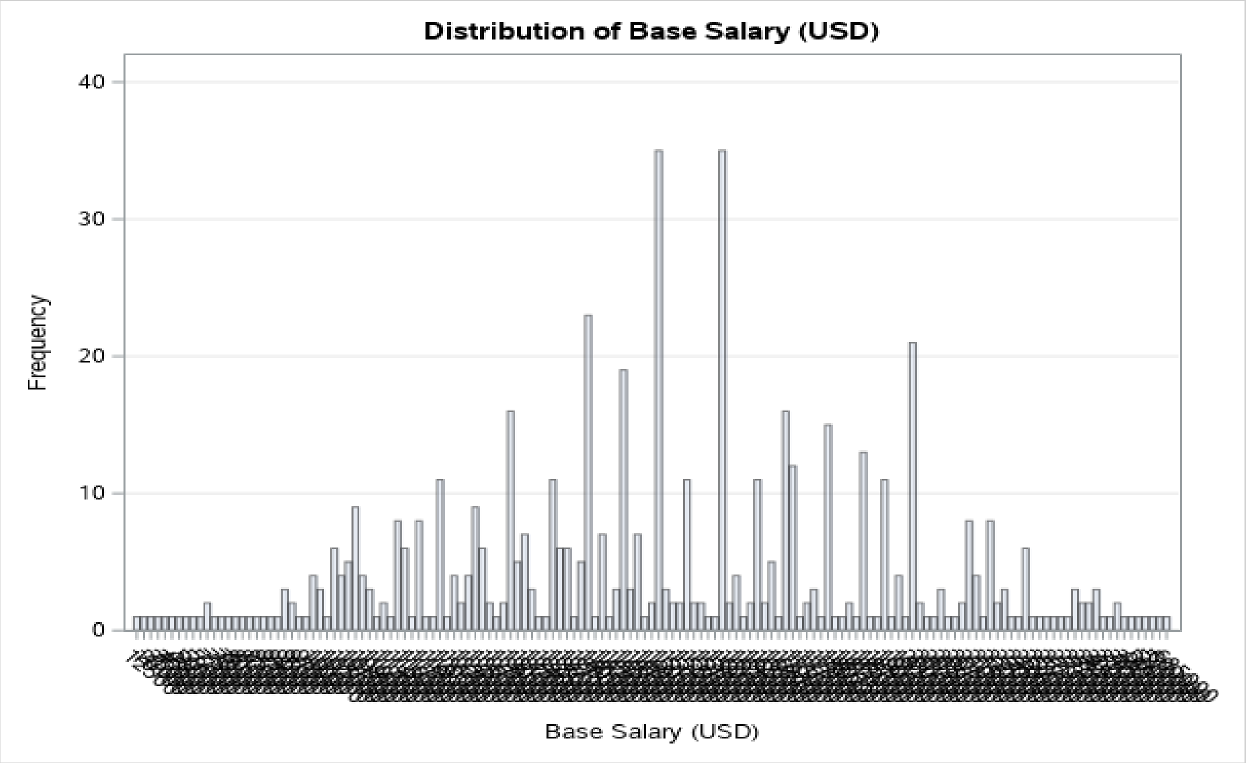


Figure 12: One-way frequency chart

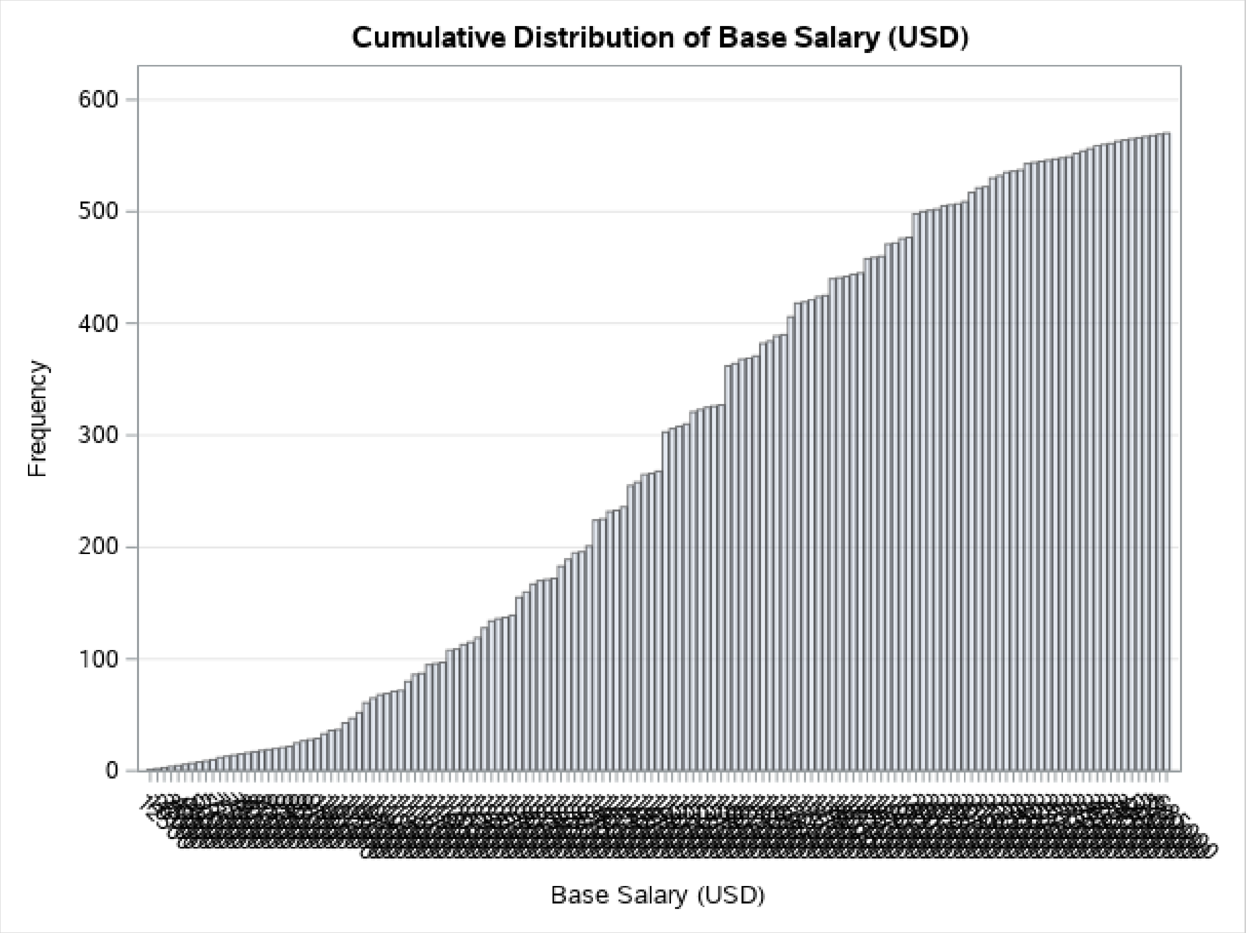


Figure 13: One-way frequency chart

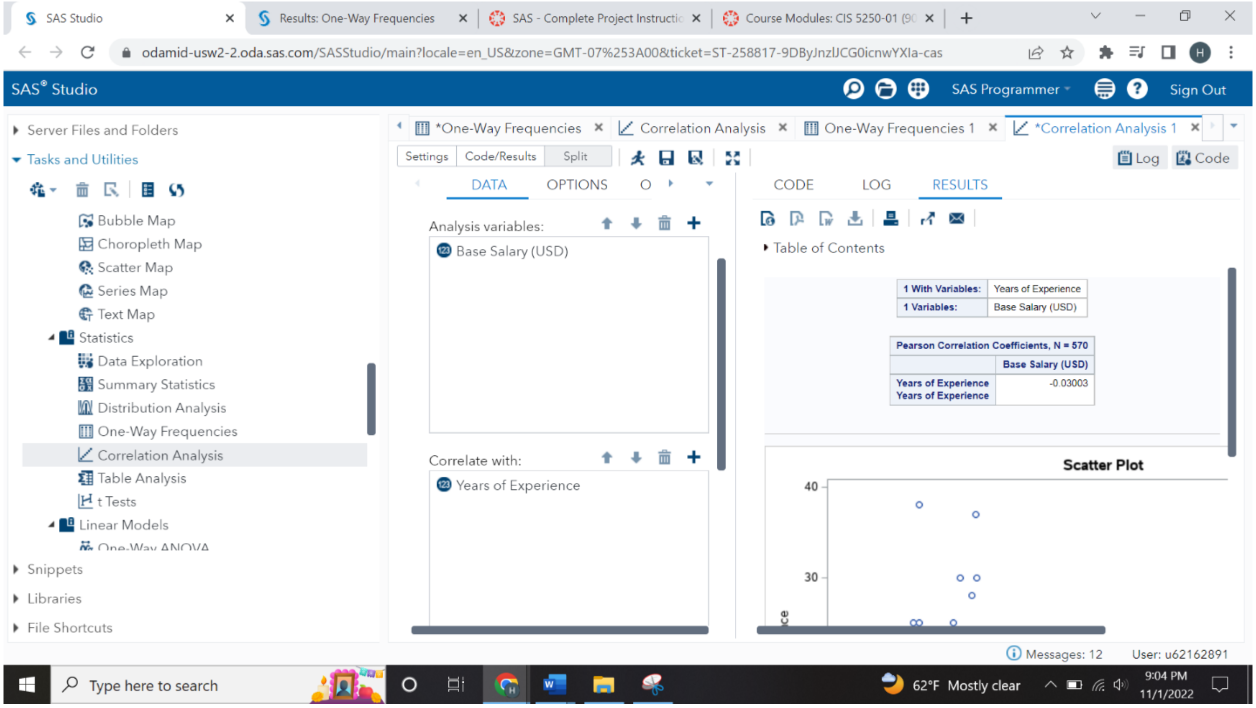
One-way frequency tables help us understand which data values are common and rare. A frequency distribution provides a visual representation of the distribution of observations within a particular test. We often use a frequency distribution table to visualize the data. These tables organize our data and are an effective way to present the results to others. Frequency tables are also known as frequency distributions because they allow you to understand the distribution of values in your dataset. The most frequently occurring values are easily identified, as are value ranges, lower and upper limits, cases that are not common, outliers, and the total number of observations between any given values.

A frequency table shows the distribution of observations based on the base salary offered by different companies. This frequency table can be used to understand which value occurs more and which value occurs less in the dataset. This table helps us know the frequency distribution and describes how frequently a value is repeated in the data set. This table shows that the base salary of 150000 USD and 160000 USD is repeated 35 times. Many values are not repeated. This table shows the variation in salary from 12500 USD to 685000 USD. We can see the distribution of the salary.

In this case, if a candidate is looking for a particular salary range, they can divide the salary category into lower, average, and high salary ranges according to their expectation and compare the salaries offered by different companies. They can see how many companies offer low, medium, and high salaries. With the help of the table, the candidate will get an idea of their salary expectation. [6]

2. Correlation Analysis

In statistics, correlation or dependence is any statistical relationship between two random variables or bivariate data, whether causal or not. Correlation analysis in research is a statistical method used to measure the strength of the linear relationship between two variables and compute their association. Correlation is a statistical measure that describes how two variables are related and indicates that as one variable changes in value, the other variable tends to change in a specific direction. Below is a correlation analysis for the variable base salary (USD) and year of experience. [5]



Analysis variables: Base Salary (USD)

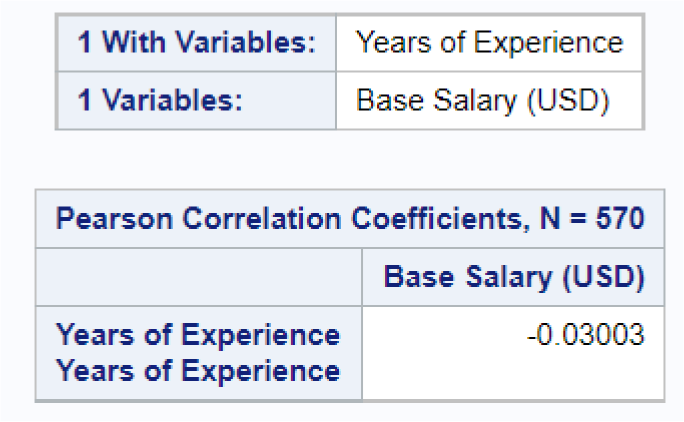
Correlate with: Years of Experience

1 With Variables:	Years of Experience
1 Variables:	Base Salary (USD)

Pearson Correlation Coefficients, N = 570	
	Base Salary (USD)
Years of Experience Years of Experience	-0.03003

Messages: 12 User: u62162891

Figure 14: Screenshot of the output



1 With Variables:	Years of Experience
1 Variables:	Base Salary (USD)

Pearson Correlation Coefficients, N = 570	
	Base Salary (USD)
Years of Experience Years of Experience	-0.03003

Figure 15: Correlation analysis table

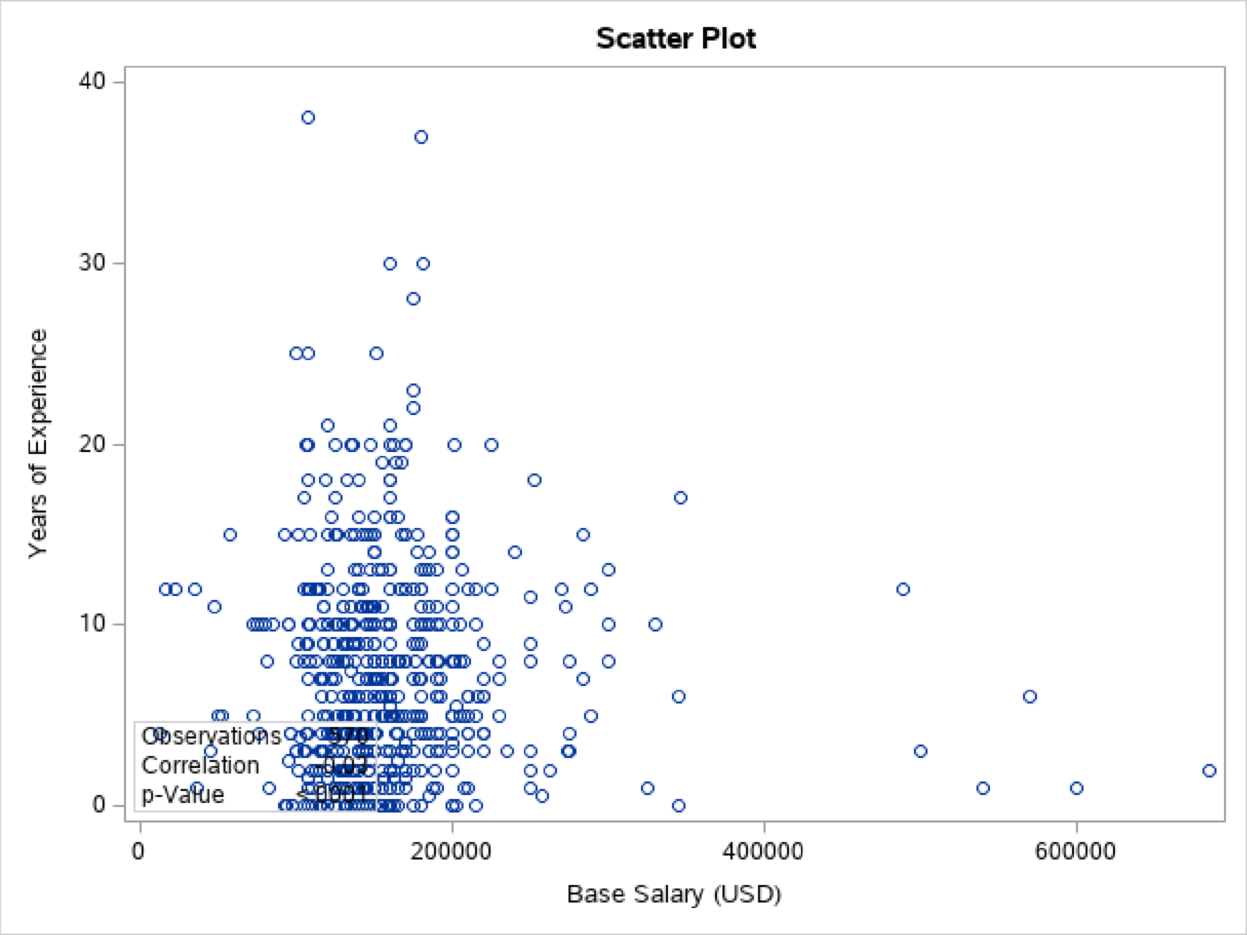


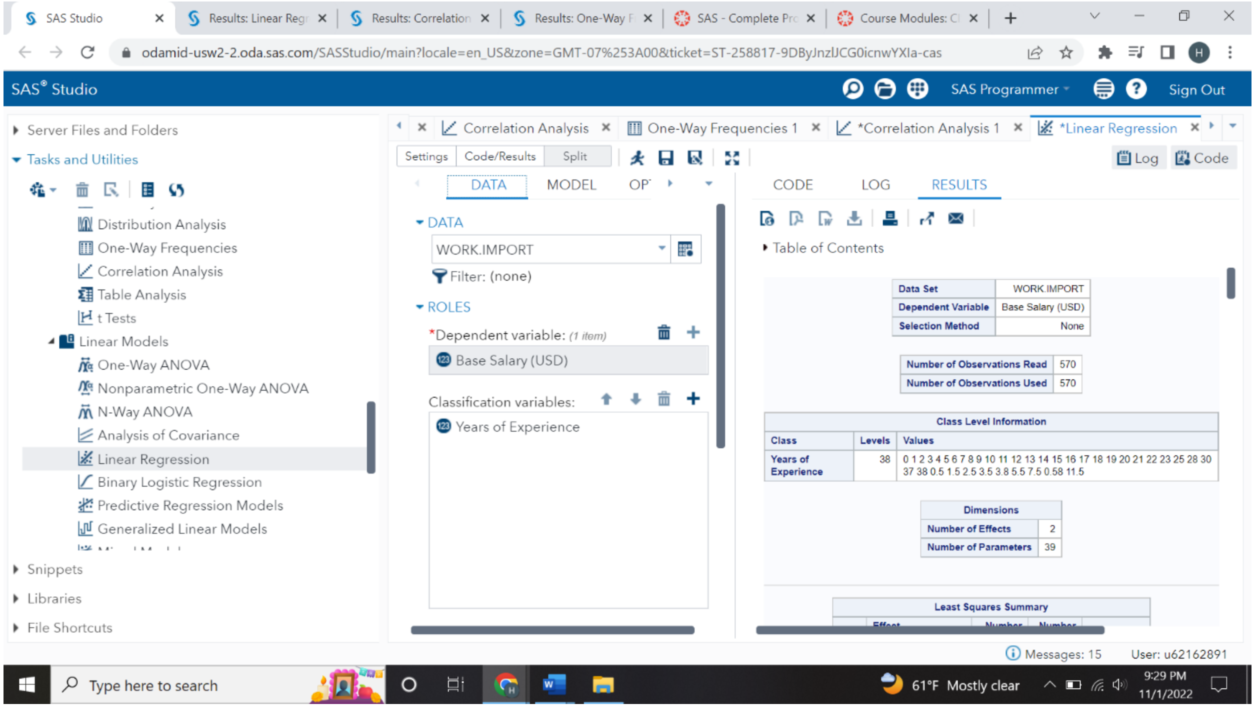
Figure 16: Correlation analysis scatter plot

In this example, the analysis variable is Base salary (USD), and the correlated variable is years of experience. Here we are checking the correlation between these two variables. We are checking if years of experience have any effect on base salary. Correlation analysis measures the strength of the linear relationship between two variables and computes their association. With the help of correlation analysis, we can calculate the level of changes in one variable due to the difference in the other. The Pearson correlation coefficient tests whether the relationship between two variables is significant. Pearson's correlation is used when you want to see if there is a linear relationship between two variables. The values range of the correlation coefficient is between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there is an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A positive correlation means the two variables move in the same direction. A negative correlation means they move in opposite directions.

The table shows the Pearson correlation coefficient as – 0.03, which shows a negative correlation. For Negative correlation, the two variables move in opposite directions, i.e., one variable increases as the other decreases, and vice versa. From this analysis, we can conclude that there is no relation between the variables, base salary, and years of experience.

3. Linear Regression

Linear Regression analysis allows you to understand the strength of relationships between variables. Using statistical measurements like R-squared / adjusted R-squared and regression analysis, we can tell you how much of the total variability in the data is explained by your model. [4]



DATA

WORK.IMPORT

Filter: (none)

ROLES

*Dependent variable: *(1 item)*

Base Salary (USD)

Classification variables:

Years of Experience

Data Set	WORK.IMPORT
Dependent Variable	Base Salary (USD)
Selection Method	None

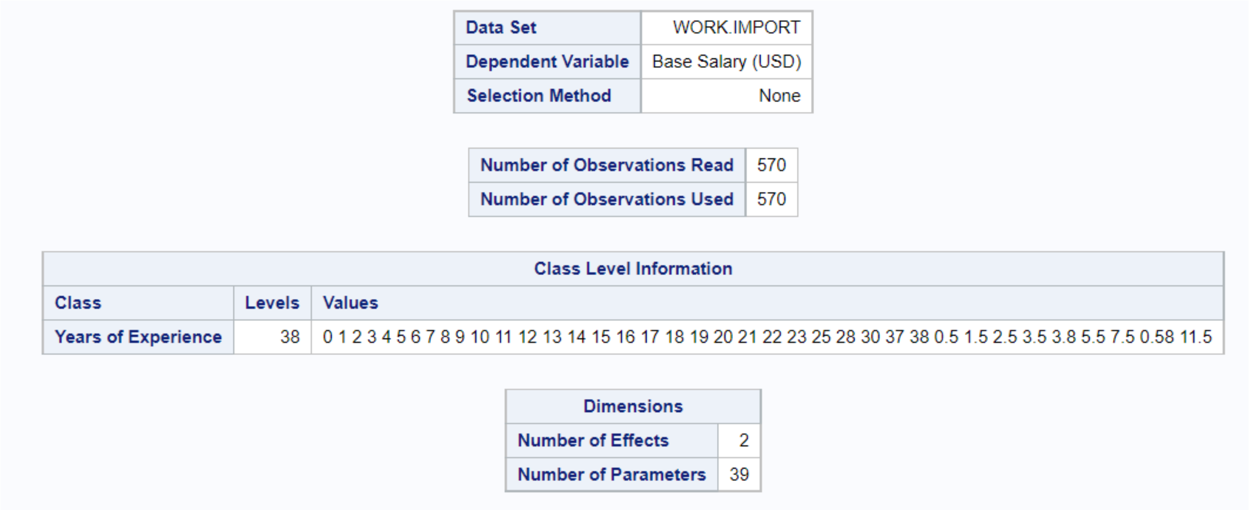
Number of Observations Read	570
Number of Observations Used	570

Class Level Information		
Class	Levels	Values
Years of Experience	38	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 28 30 37 38 0.5 1.5 2.5 3.5 3.8 5.5 7.5 0.58 11.5

Dimensions	
Number of Effects	2
Number of Parameters	39

Least Squares Summary

Figure 17: Screenshot of the output

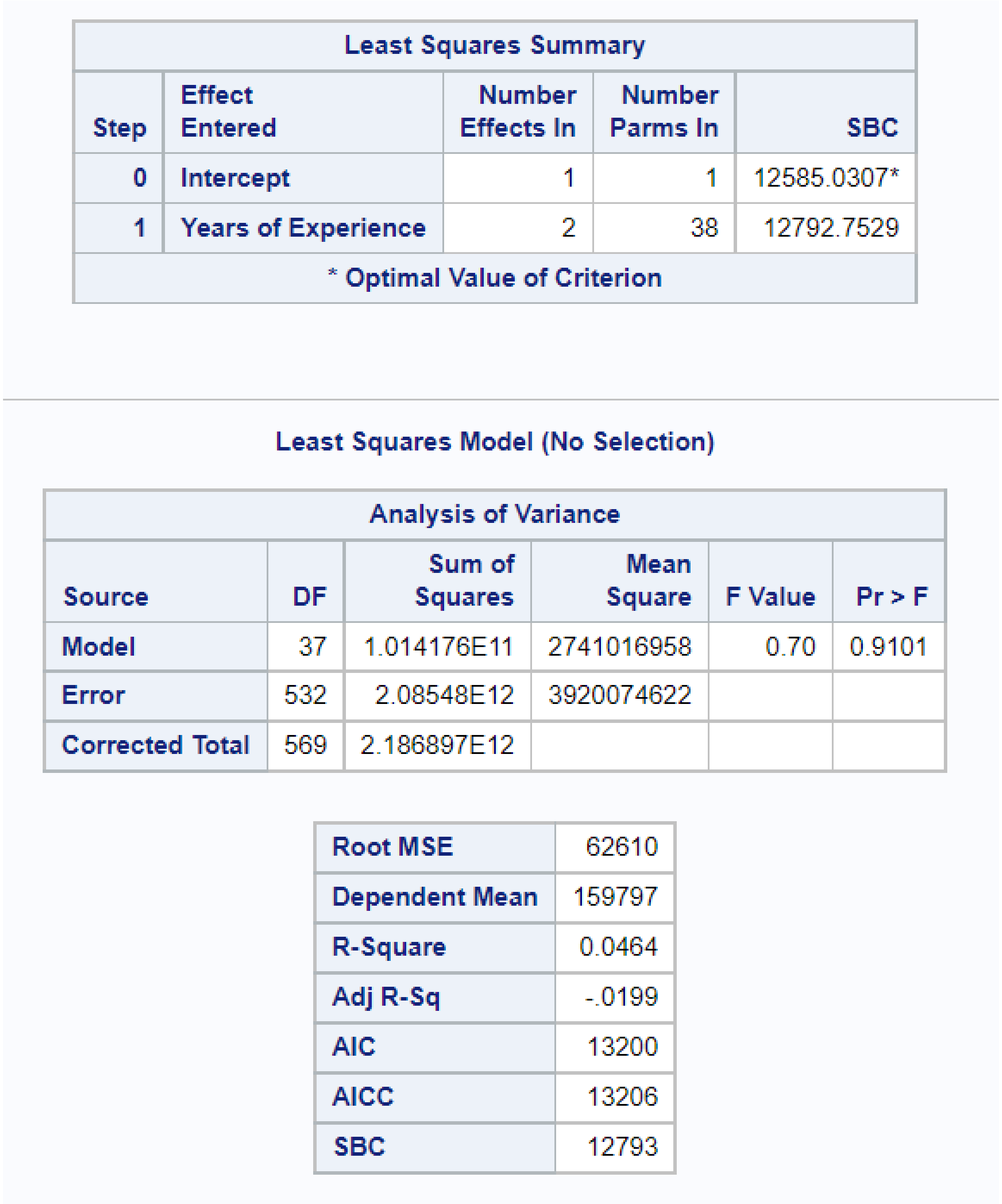


Data Set	WORK.IMPORT
Dependent Variable	Base Salary (USD)
Selection Method	None

Number of Observations Read	570
Number of Observations Used	570

Class Level Information		
Class	Levels	Values
Years of Experience	38	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 28 30 37 38 0.5 1.5 2.5 3.5 3.8 5.5 7.5 0.58 11.5

Dimensions	
Number of Effects	2
Number of Parameters	39



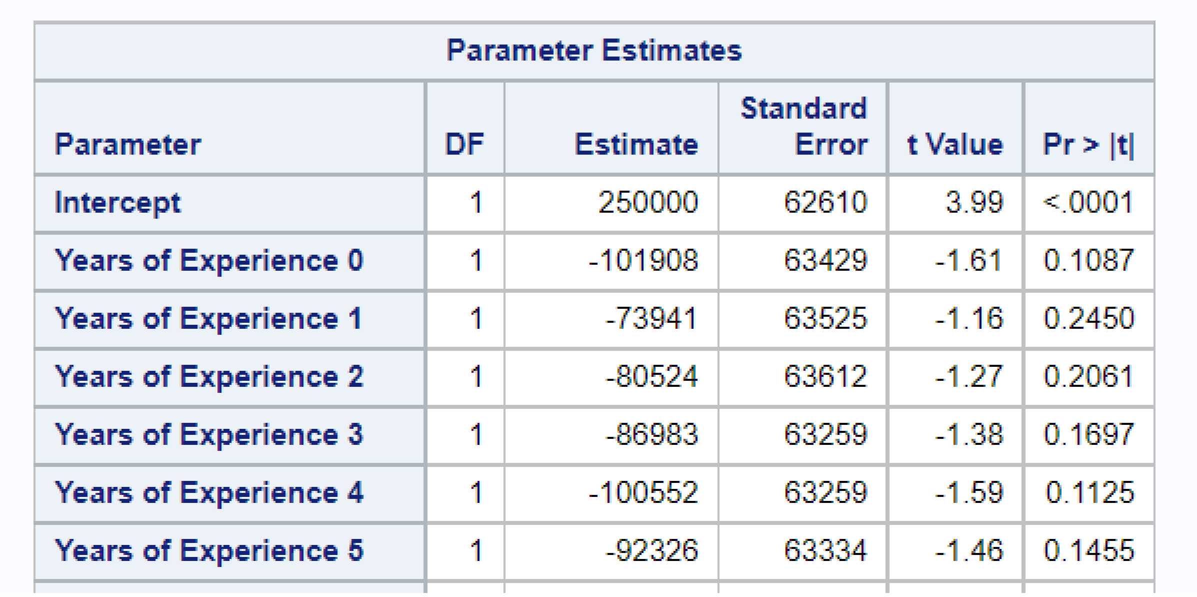
Least Squares Summary				
Step	Effect Entered	Number Effects In	Number Parms In	SBC
0	Intercept	1	1	12585.0307*
1	Years of Experience	2	38	12792.7529

* Optimal Value of Criterion

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	37	1.014176E11	2741016958	0.70	0.9101
Error	532	2.08548E12	3920074622		
Corrected Total	569	2.186897E12			

Root MSE	62610
Dependent Mean	159797
R-Square	0.0464
Adj R-Sq	-.0199
AIC	13200
AICC	13206
SBC	12793



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	250000	62610	3.99	<.0001
Years of Experience 0	1	-101908	63429	-1.61	0.1087
Years of Experience 1	1	-73941	63525	-1.16	0.2450
Years of Experience 2	1	-80524	63612	-1.27	0.2061
Years of Experience 3	1	-86983	63259	-1.38	0.1697
Years of Experience 4	1	-100552	63259	-1.59	0.1125
Years of Experience 5	1	-92326	63334	-1.46	0.1455

Figure 18: Linear Regression analysis table

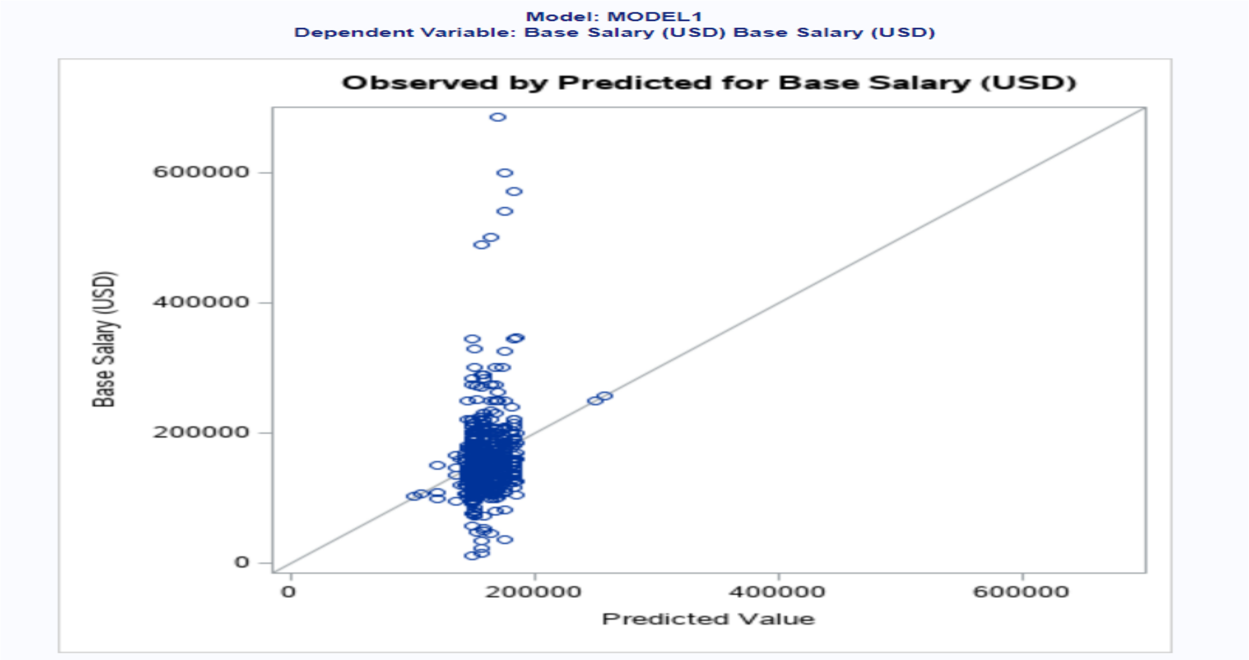


Figure 19: Linear Regression analysis chart

Linear regression analysis predicts a variable's value based on another variable's value. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. In our example, the dependent variable is base salary, and the independent variable is the year of experience.

Linear Regression Equation

$$Y = mx + c$$

In the above equation, m is the slope, c is the constant, and y is the y-intercept

From the linear regression table, we can write the liner regression equation as

$$Y = 1 * (year\ of\ experience) + 250000$$

R square value is approximately 0.0464, which is 4.64%. This shows a correlation between base salary and years of experience. This indicates very low correlations. There were 570 observations. No missing values are reported. The r-squared value is 4.64%, which is the correlation between base salary and years of experience in this example.

For p=0.5, it is statistically significant because it is greater than 0.0001, shown in the table. This means that any variation in the base salary can be explained by the change in the year of experience. For p=0.05, the results are statistically significant because it is greater than 0.0001; the model value is given in the table. We can conclude that years of experience explain the variation in base salary has significantly less impact on this variation.

Reference

1. Monnappa, A. (2022, March 29). Why data science matters and how it powers business in 2022. Simplilearn.com. Retrieved October 6, 2022, from https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article
2. Morgan, L. (2022, July 15). The Future of Data Science: Career Outlook and Industry Trends. SearchEnterpriseAI. Retrieved October 6, 2022, from https://www.techtarget.com/searchenterpriseai/feature/The-future-of-data-science-jobs
3. SinHacker. (2019, May 27). When job hunting meets Data Science (Part 1). Medium. Retrieved October 6, 2022, from https://towardsdatascience.com/when-job-hunting-meetsdata-science-part-1-e8f64867d8c
4. About linear regression. IBM. (n.d.). Retrieved November 2, 2022, from https://www.ibm.com/topics/linear-regression
5. Everything you need to know about interpreting correlations. (n.d.). Retrieved November 3, 2022, from https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8
6. Foundation, C. K.-12. (n.d.). 12 foundation. CK. Retrieved November 2, 2022, from https://www.ck12.org/book/ck-12-middle-school-math-concepts-grade-8/r14/section/1.2/