Data Visualization

Introduction:

Data Visualization is a visual representation and presentation of a data to facilitate the understanding of the intended users (Andy, 2019). Visual representation is all about the selection of right charts to display the features of relevant data. The term presentation involves the choice of design for the charts such as the interactivity, usage of colours and the features of annotation (Andy, 2019). Before applying visualization on any kind of data, it is crucial to carry out the cleaning process of data. Data cleaning or pre-processing is important as it ensures improvement in the quality of data which directly impacts the data analysis. The report illustrates about the various data cleaning approaches and visualization techniques applied on a dataset called ‘data scientist.csv’ to derive valuable insights from it.

Dataset Details:

The dataset consists of details of various jobs in United States and it includes 3909 tuples or rows and 16 attributes or columns. The attribute ‘index’ contains a unique numeric value for each record in the dataset. It includes details for different job titles such as data scientist, data analyst, data engineer, etc. In addition to the sixteen attributes few additional features have been extracted to carry out a better analysis process. These features include minimum and maximum pay of each job title from the salary range in the salary column, minimum and maximum revenue from the range of revenues in the revenue attribute.

Data Cleaning:

<u>Mode Substitution:</u> This process involves identifying the rows which does not have proper value and cannot be used for analysis and replacing it with the most frequent values in the column. Such improper values include but not limited to ‘unknown’, not applicable’, etc. Replacing the null values than removing the rows with null values ensure there is no loss or drop of information even though there is an opportunity to lose the integrity of the data (Tableau). It is easy to implement and understand and is fast as well irrespective of data size (TDS, (2021). The distribution of data might get impacted slightly or even heavily depends on the number of missing values (TDS, (2021).

For Column Headquarters;

```
 df['Headquarters'] = df['Headquarters'].replace(-1, df['Headquarters'].mode()[0]
```

For Column Size;

```
df['Size'] = df['Size'].replace('-1', df['Size'].mode()[0])

df['Size'] = df['Size'].replace('Unknown', df['Size'].mode()[0])
```

For Column ‘Type of Ownership’;

```
df['Type of ownership'] = df['Type of ownership'].replace('Unknown', df['Type of ownership'].mode()[0])

df['Type of ownership'] = df['Type of ownership'].replace('-1', df['Type of ownership'].mode()[0])
```

For Column Industry;

```
def most_common(lst):
    return max(set(lst), key=lst.count)
industrylist=[]
def removenull(val):
    if(val == '-1'):
        print(' it is -1')


    else:
        industrylist.append(val)
    return val



df['Industry'].apply(lambda x: removenull(x))
df['Industry'] = df['Industry'].replace('-1', mode(industrylist))


industrylist.clear()
df['Sector'].apply(lambda x: removenull(x))
df['Sector'] = df['Sector'].replace('-1', mode(industrylist))


industrylist.clear()
df['Competitors'].apply(lambda x: removenull(x))
df['Competitors'] = df['Competitors'].replace('-1', mode(industrylist))
```

Mean Substitution: This technique is all about finding the rows which does not have proper values and cannot be used for analysis and then replacing it with the average of the proper values in the same column. This process has been applied for the column ‘Rating’. The below code has been used to substitute the column with mean value

```
df['Rating'] = df['Rating'].replace(-1.0, df['Rating'].mean())
```

Median Substitution: This technique has been used to replace the null values of the column ‘Founded’ with the centre or median value of the same column. As the column ‘Founded’ contains the values as years which exhibit the feature of skewness, median substitution is the right choice of pre-processing than mean or mode substitution (Analytics Vidhya, 2020). The below code has been used to substitute the null values with the median value.

```
df['Founded'].median()

df['Founded'] = df['Founded'].replace(-1, df['Founded'].median())
```

Additional Features: Since the columns ‘ Salary’ and ‘Revenue’ contains the range of values instead of having a single value, new columns such as ‘Min Salary’, ‘Max Salary’, ‘Min Revenue’ and ‘Max Revenue’ has been created. This has been achieved by splitting the range of values in the respective columns and then extracting those values.

Variation Reduction: The column ‘Job Title’ has so many different extensions for the job titles such as data scientist, data analyst, business scientist, etc. Having such higher number of different titles could result in redundant analysis. Hence, the job titles with variations have been reduced to some of the commonly known job titles. The below python code has been used to reduce the variations in job titles.

```
def main_job_title(title):

    if( 'Senior' in title):

        return 'Senior Scientist'

    if( 'Associate' in title or 'Assistant' in title):

        return 'Associate Scientist'

    if( 'Data Scientist' in title or 'Data Science' in title or 'Data Scientists' in title):

        return 'Data Scientist'

    if( 'Data Engineer' in title or 'DATA ENGINEER' in title):

        return 'Data Engineer'

    if( 'Business Intelligence' in title or 'Business Data' in title  or 'Business Analytics' in title):

        return 'Business Analyst'
    if( 'Data Analyst' in title or 'Data & Analytics' in title or 'Data Analytics' in title):
        return 'Data Analyst'
    if( 'Data Modeler' in title):
        return 'Data Modeler'
    if( 'Data Architect' in title):
        return 'Data Architect'
    if( 'AI/ML' in title or 'AI' in title or 'Machine Learning' in title):
        return 'AI/ML Engineer'
    if( 'Clinical' in title or 'CLINICAL' in title):
        return 'Clinical Scientist'
    if( 'Computer Scientist' in title or 'COMPUTER SCIENTIST' in title):
        return 'Computer Scientist'
    if( 'Medical Lab' in title):
        return 'Medical Lab Scientist'
    if( 'Environmental' in title):
        return 'Environmental Scientist'
    if( 'Doctoral' in title or 'Research Scientist'):
        return 'Research Scientist'
    if( 'Food' in title):
        return 'Food Scientist'
    if( 'Research' in title or 'RESEARCH' in title):
        return 'Other Research Roles'
    if( 'Engineer' in title or 'Developer' in title):
        return 'Software Engineer/Developer'
    if( 'manager' in title or 'director'):
        return 'Managerial Roles'
    else:
        return 'Other Scientist Roles'


df['Job Title']=df['Job Title'].apply(lambda x: main_job_title(x))
```

Column Removal: The attribute ‘Easy Apply’ contains null values which accounted for more than 75% of the total number of records. Hence it is wise to remove that attribute as it will not create a big impact in the analysis of data.

```
#Checking for 75% Invalid Values in 'Easy Apply'

Invalid_values = ['-1','NaN','Unknown',' ']

df1 = pd.read_csv('Initial_cleaned.csv',na_values=Invalid_values)

Invalid_count_easy_apply=df1['Easy Apply'].isnull().sum()

rows = df1['Easy Apply'].count()

percentage_Invalid = (Invalid_count_easy_apply*100)/rows

print('percentage Invalid Values = ', percentage_Invalid)


#Removing the Column 'Easy Apply'

if percentage_Invalid >=75.0:

   df1.drop('Easy Apply',axis=1,inplace=True)
```

Selection of Visualization:

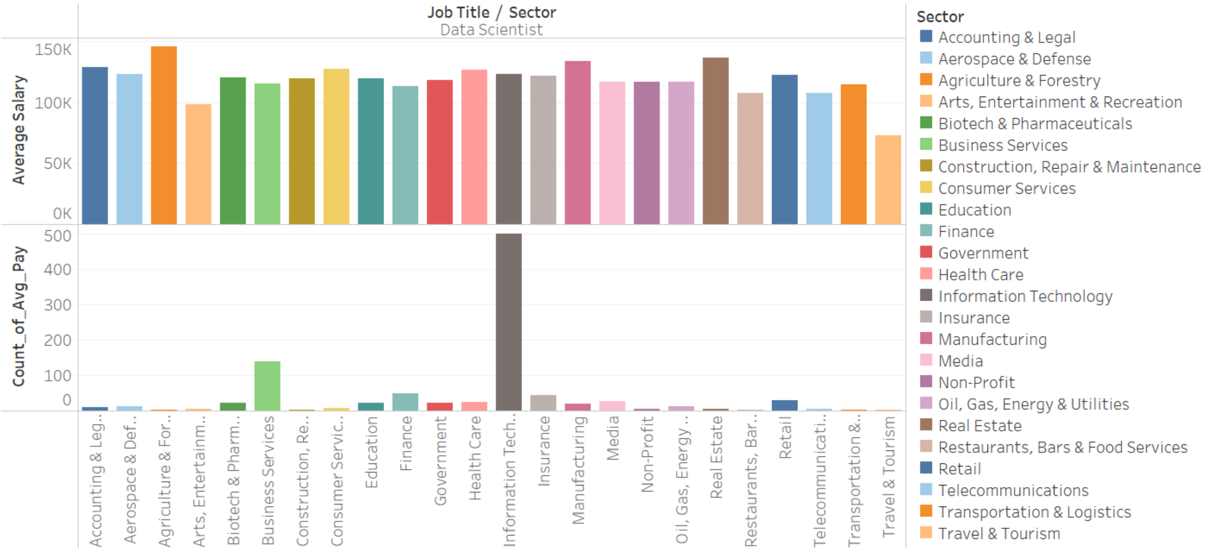
The dataset contains both numeric and text data hence it is appropriate to consider the visualizations which can support both type of data. The famous ‘Word Cloud’ is more appropriate for the text data thought the dataset contains text it does not have any data such as user inputs or comments. For example, the column ‘Rating’ contains a numeric value of user input but does not have any column with text data corresponding to the user rating. Hence, text visualization techniques such as Word Cloud are not the right choices for this dataset and can be ignored. The right selection choice can be between charts, graphs and maps. Graphs are more suitable for the time-oriented data to visualize the progress of data, however, as the dataset does not contain any time-related data, graphs visualizations can be ignored. The dataset contains details such as sector, industry, type of ownership, location, company name, etc. which can be used to display in the hierarchical format of visualization. Hence, the techniques which are suitable for hierarchical data such as Treemap are more appropriate for this dataset.

Visualization:

Best Sector for Data Scientist:

The side-by-side bar chart identifies the sector which provides higher salary for the data scientist role. The average of maximum salary has been taken and plotted against each sector. It is evident from the chart that the highest average of maximum salary comes from the real estate sector, however there is not huge difference with the next highest paying sectors such as manufacturing. However, the value of averages could have been based on the number of jobs in that sector as well. Hence it is important to find the number of opportunities in each sector for data scientist to make the final decision on the best sector for data scientist career. Again, side by side chart has been used to identify the number of jobs for data scientist in each sector. It is clearly evident that the Information technology sector wins the race by a hug margin. Though sectors such as real estate offers highest pay for data science roles, the best sector based on the combination of number of opportunities and the average highest pay is Information technology for a data scientist.

Avg. Pay & Job Count - Data Scientist

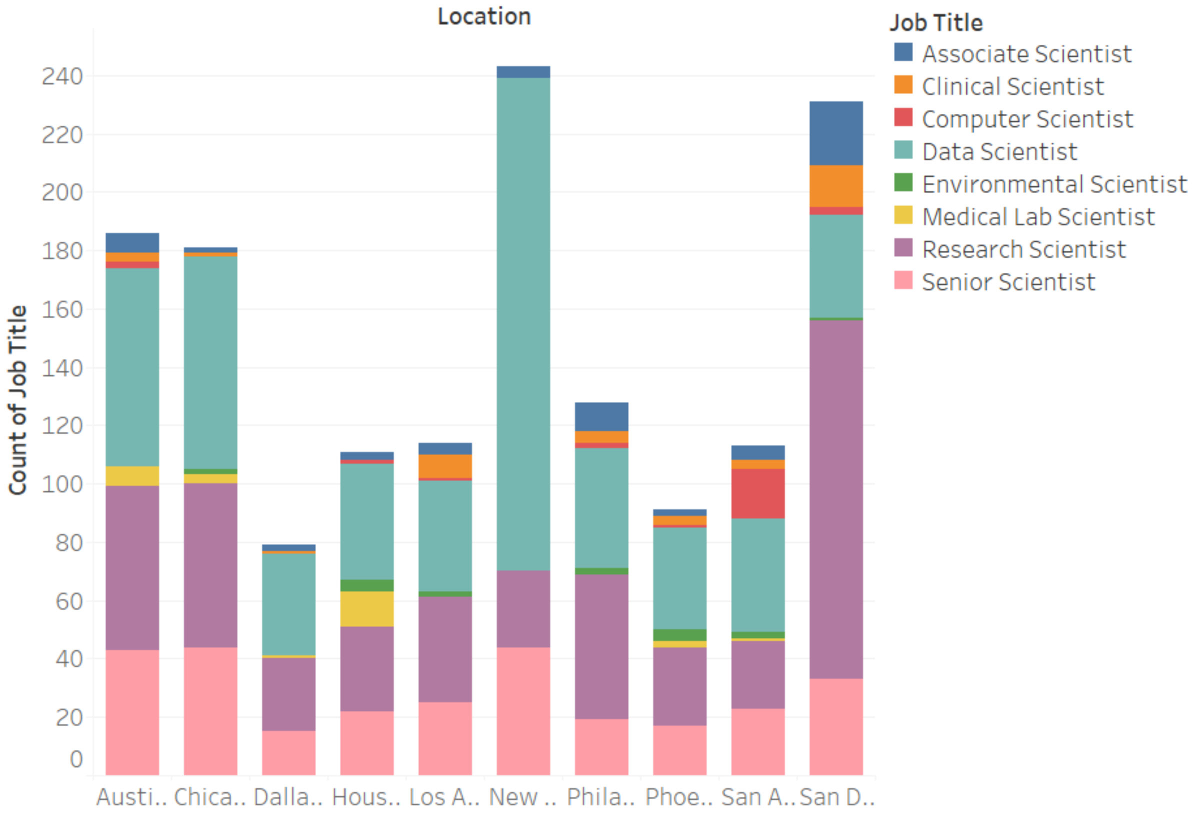


Average Salary and Count_of_Avg_Pay for each Sector broken down by Job Title. Color shows details about Sector. The view is filtered on Job Title, which keeps Data Scientist.

Distribution of Scientist Jobs:

The stacked bar chart provides information about the number of different scientist jobs across the locations in the United States. Stacked bar chart is a type of chart that displays multiple data points on top of each other and helps to illustrate how each points relate to the total (Indeed, 2021). It can be inferred from the chart that the roles 'Data Scientist' and 'Research Scientist' are two of the hot jobs among all followed by the senior scientist role. The remaining scientist roles have very little opportunities in all the locations. The city New York holds the most number of data scientist jobs while for the research scientist role it is the city San Diego which holds more jobs.

Sheet 1

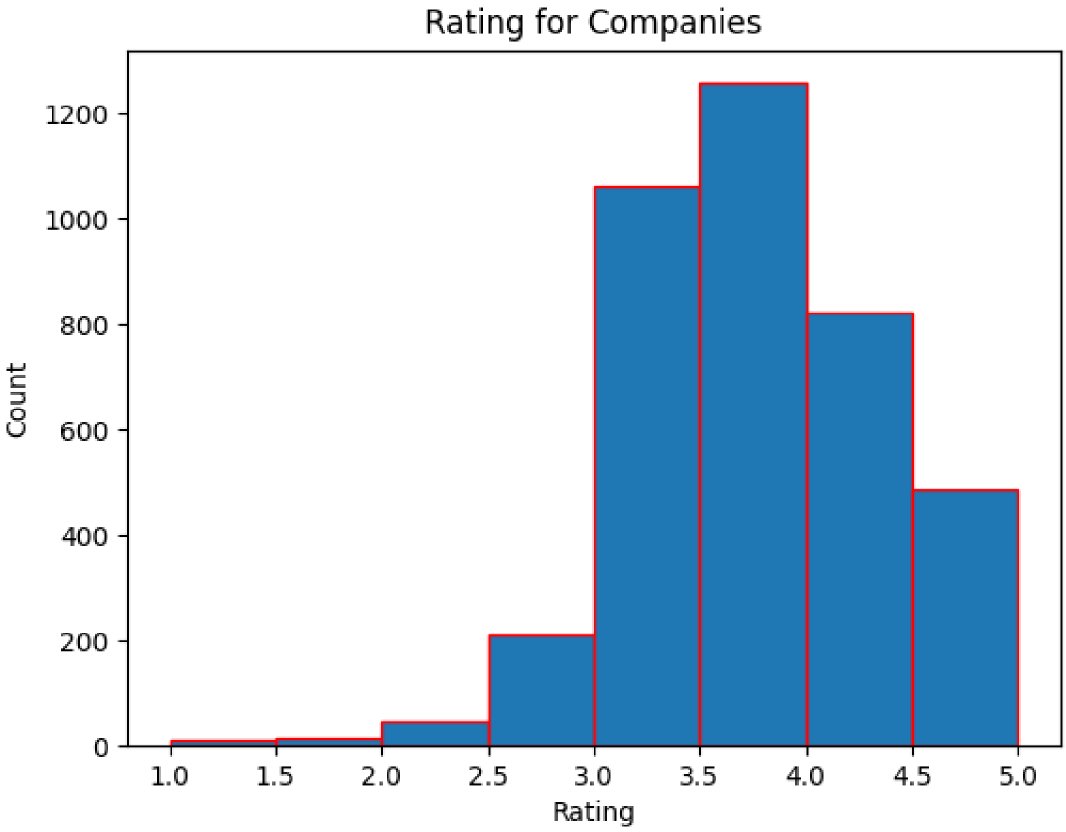


Count of Job Title for each Location. Color shows details about Job Title. The view is filtered on Job Title and Location. The Job Title filter has multiple members selected. The Location filter has multiple members selected.

Rating Distribution:

The histogram chart identifies the total number of different ratings for the companies in the United States. Histogram are an excellent way to link arithmetical data as it provides user-friendly and understandable format to see where most of the values fall in a measurement scale (Latest Quality, 2017). It is clearly evident from the chart that there are not many companies with poor ratings while the companies with high ratings accounts for a decent number. Most of the ratings occur between 3.5 to 4.0 which states that companies are doing pretty good job with respect to many factors if not all. The below code has been used for the visualization of histogram.

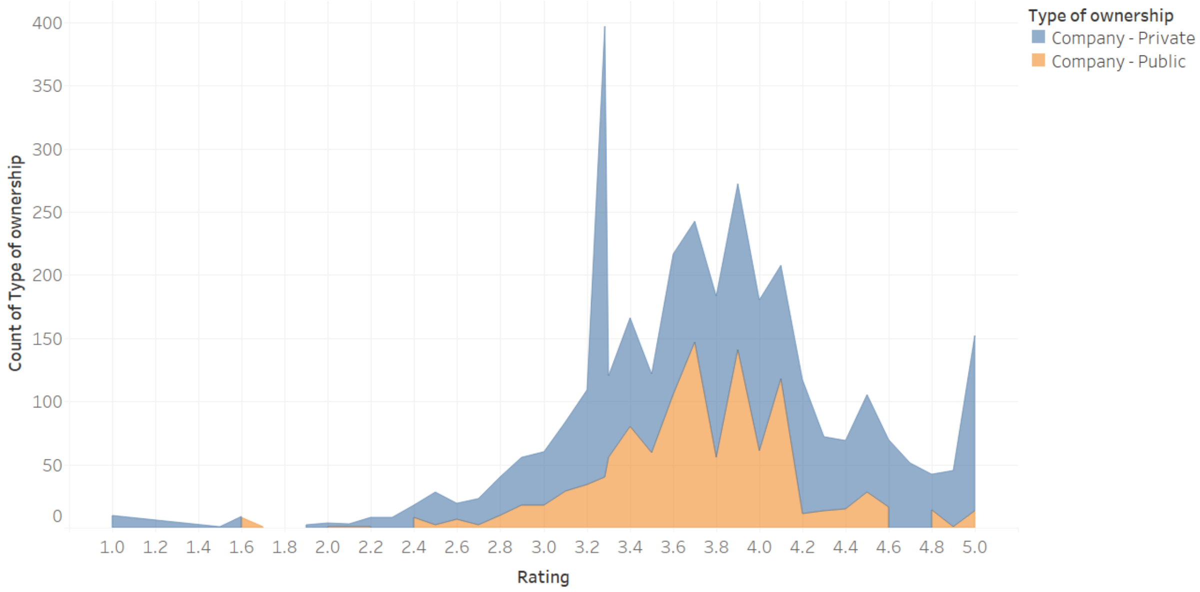
```
import matplotlib.pyplot as plt

x=df2['Rating']

plt.hist(x,[1,1.5,2,2.5,3,3.5,4,4.5,5],ec='red')

plt.xlabel("Rating")

plt.ylabel("Count")
plt.title("Rating for Companies")
plt.show()
```



Public vs Private Firm:

The area chart illustrates the distribution of the ratings for public and private companies. Most of the ratings for both public and private companies lie between the ratings 3.0 and 4.0. There are few exceptions with very poor ratings of less than 2.0 which is in particular for few private companies which holds the least ratings. It can be inferred that working in both public and private firms with respect to the data scientist jobs does not have much difference. Area chart has been used for this information visualization as it shows comparisons between different groups and makes research findings more readable and scannable (Indeed, 2021). The chart offers the facility to understand the data points without the need to interpret numerical information (Indeed, 2021).

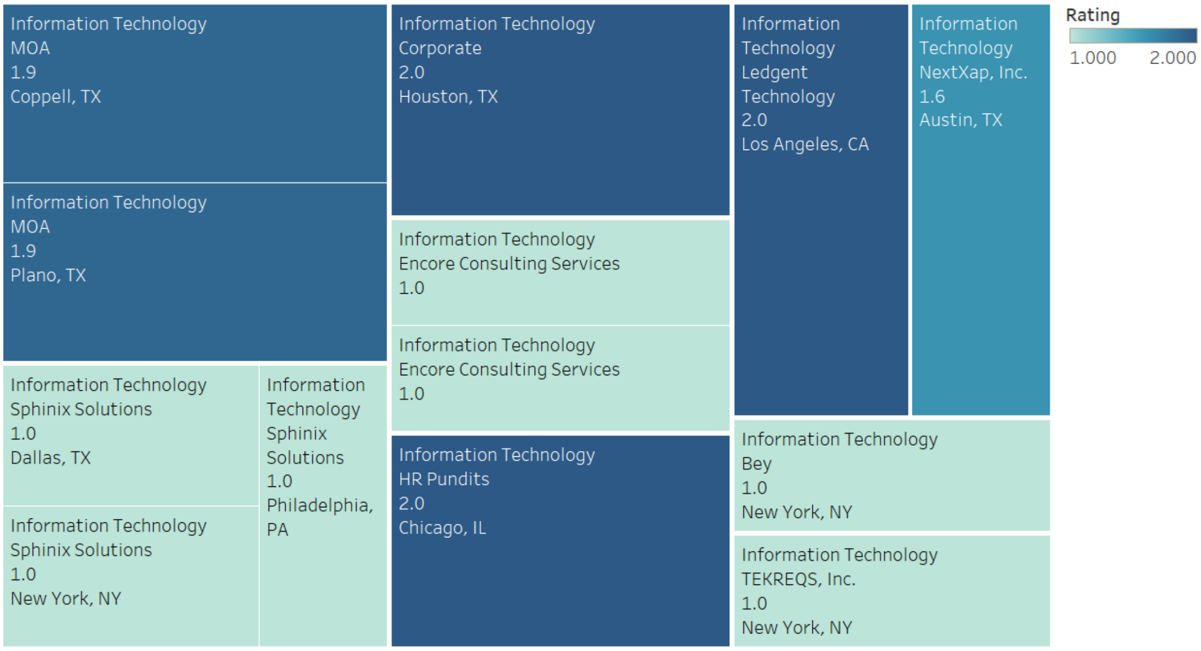
Public vs Private Firm



The plot of count of Type of ownership for Rating. Color shows details about Type of ownership. The view is filtered on Type of ownership, which keeps Company - Private and Company - Public.

Poor IT Companies:

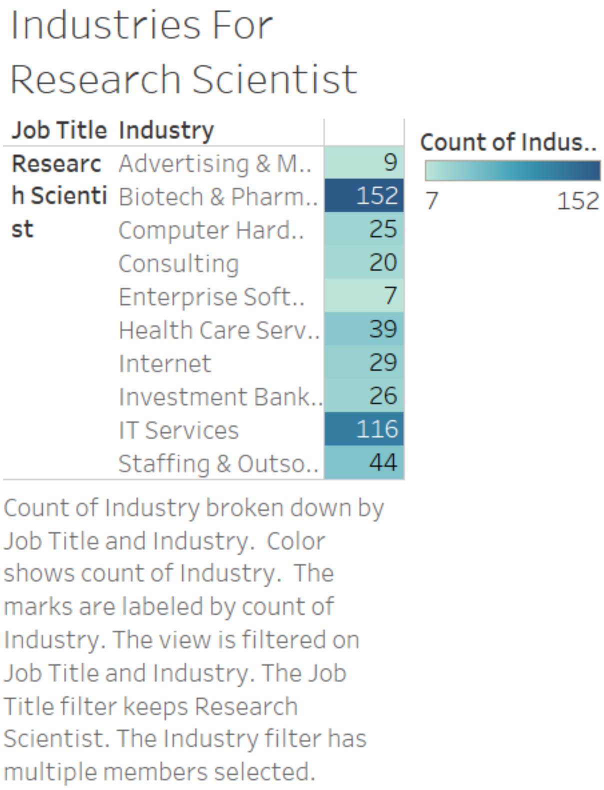
Companies with Low Ratings in IT



Sector, Company Name and Location. Color shows sum of Rating. Size shows sum of Rating. The marks are labeled by Sector, Company Name and Location. The view is filtered on Sector, Location, sum of Rating and Company Name. The Sector filter keeps Information Technology. The Location filter keeps 10 of 191 members. The sum of Rating filter ranges from 1.000 to 2.000. The Company Name filter keeps no members.

The Treemap plot has been used to provide information about the list of companies with low ratings in the information and technology sector. Treemap visualization facilitates users to see large hierarchically structured information spaces. Hierarchical information used to contain two types of information namely structural information associated with hierarchy and the content associated with each node. Tree-maps can be able to depict both structure and content type of information (Brian and Ben, 1991). It offers efficient utilization of space, interactivity and comprehensive method of presentation (Brian and Ben, 1991).

Industry For Research:



Industries For Research Scientist

Job Title	Industry	
Research Scientist	Advertising & M..	9
	Biotech & Pharm..	152
	Computer Hard..	25
	Consulting	20
	Enterprise Soft..	7
	Health Care Serv..	39
	Internet	29
	Investment Bank..	26
	IT Services	116
	Staffing & Outso..	44

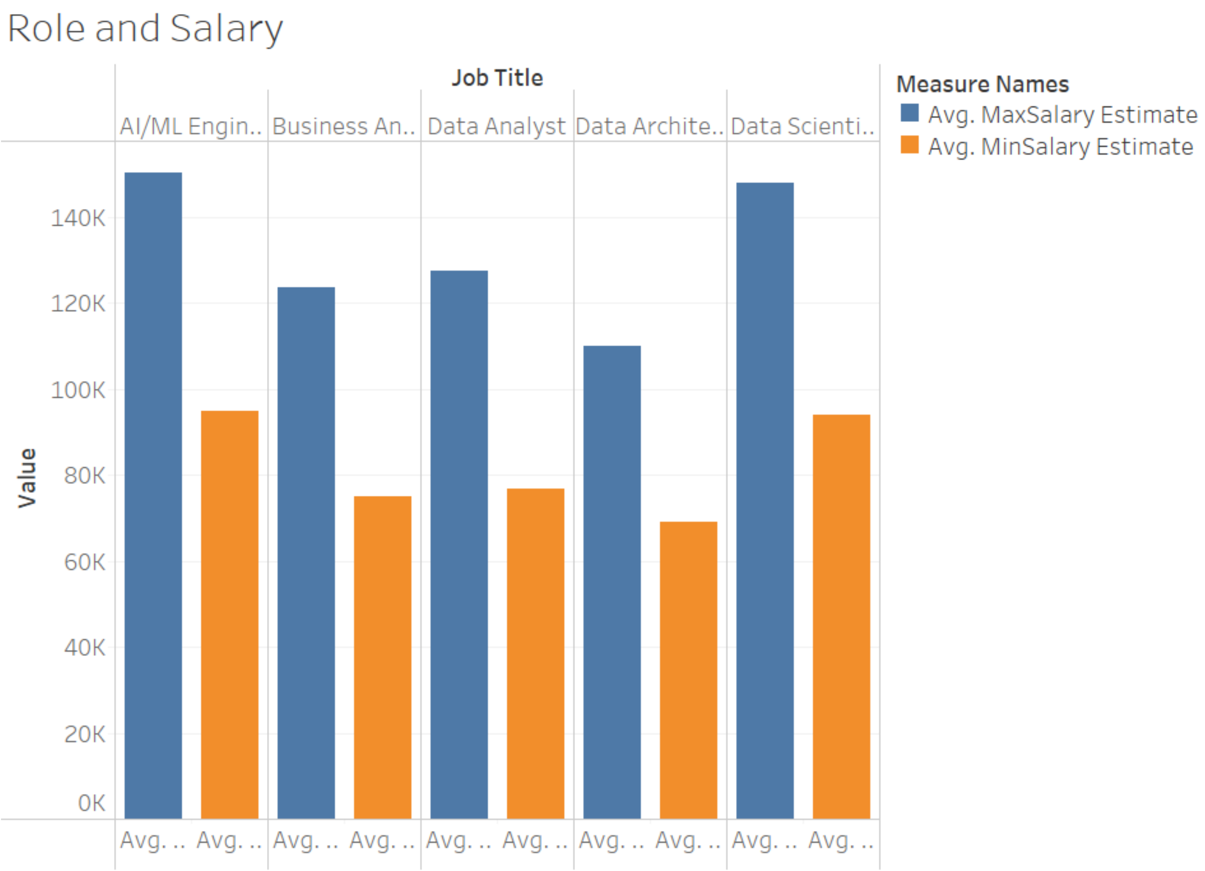
Count of Indus..
7
152

Count of Industry broken down by Job Title and Industry. Color shows count of Industry. The marks are labeled by count of Industry. The view is filtered on Job Title and Industry. The Job Title filter keeps Research Scientist. The Industry filter has multiple members selected.

The chart ‘highlight tables’ illustrates the top ten industries based on the availability of number of research jobs. From the table, it is clear that the industry ‘Biotech and Pharmeacutical’ holds the maximum number of research scientist positions followed by the IT industry. The other industries such as college and university, health care services, staffing and outsourcing and investment banking offers decent number of research scientist jobs. It can be inferred from the data that best industries for research oriented jobs are Biotech and IT services.

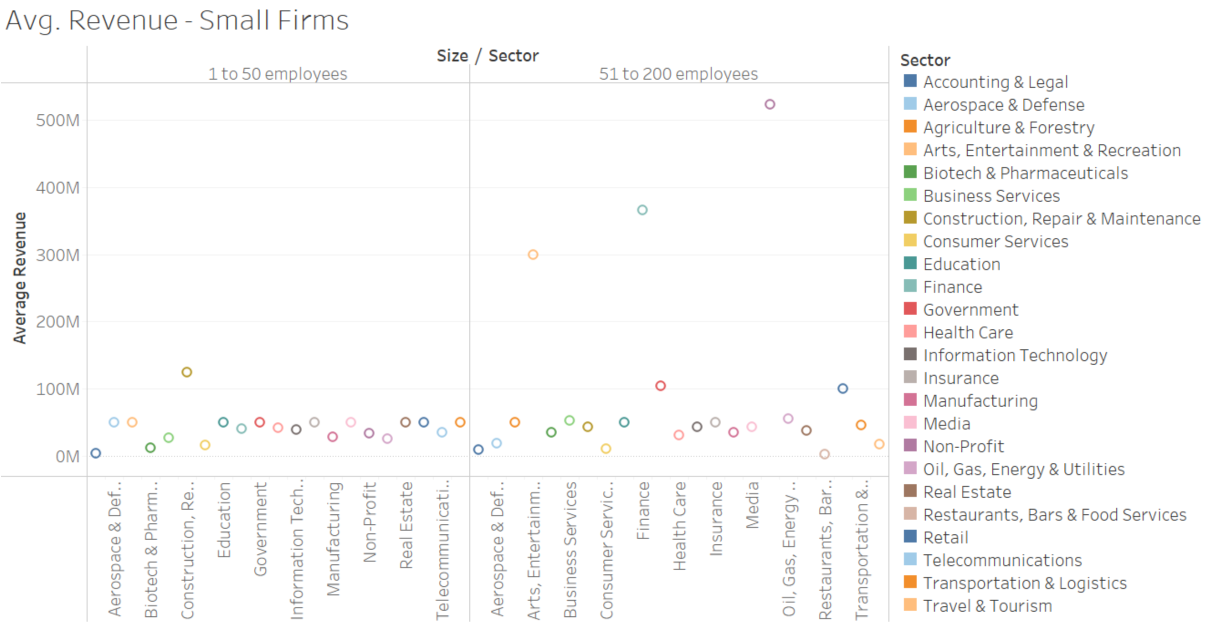
Most Lucrative Role:

The side-by-side bar chart identifies the most lucrative job using the average of maximum and minimum salary for some of the crucial job titles. It is evident the most sought-after jobs could be machine learning Engineer (AI/ML) and data scientist. The roles business analyst and data analyst are more or less similar while the least paying job is data architect as compared to the other four jobs.



Avg. MaxSalary Estimate and Avg. MinSalary Estimate for each Job Title. Color shows details about Avg. MaxSalary Estimate and Avg. MinSalary Estimate. The view is filtered on Job Title, which keeps AI/ML Engineer, Business Analyst, Data Analyst, Data Architect and Data Scientist.

Small Sized Companies:



Average Revenue for each Sector broken down by Size. Color shows details about Sector. The view is filtered on Size, which keeps 1 to 50 employees and 51 to 200 employees.

The side-by-side circle chart depicts the sectors where there is good scope for small size companies to generate revenues. This chart type enables to add greater number of dimensions or measures and is effective in pointing out the outliers (The Information Lab, 2015). Interestingly, the non-profit sector holds best average revenue measurement followed by Finance and Arts sectors for company sizes range from 51 to 200. In the case of companies with size from 1 to 50, construction sector creates good revenues as compared to other sectors. The visualization has been created using the data visualization tool called tableau.

Decision Making:

The different visualizations carried out used to make some key decisions makings. The decisions include but not limited to selecting the right sectors for the start-up companies with respect to revenue generation, identifying the best sector for the role of data scientist, finding the best industries for the research-based roles and locating the ratings of the companies in particular picking the companies which have poor ratings. The average revenues for the small-sized companies helps to identify the scope for a start-up company to invest in the particular sector. Number of research-based positions across the industries helps to find the scope and demand of research jobs in different industries while the distributions related to ratings such as public versus private firms and companies with low ratings helps in deciding about the good companies.

References

Andy, K. (2019). Data Visualization. London. SAGE Publications Ltd.

Brian, J. and Ben, S. (1991). Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures, *Proceeding Visualization '91,* p.284-291

Indeed, (2021). A Guide to Area Charts. Available from:

https://www.indeed.com/career-advice/career-development/a-guide-to-area-charts [Accessed May 12 2022]

Indeed, (2021). A Guide to Stacked Bar Charts. Available from:

https://www.indeed.com/career-advice/career-development/a-guide-to-stacked-bar-charts [Accessed May 13 2022]

Analytics Vidhya, (2020). Handling Missing Data: Data pre=processing. Available from:

https://medium.com/analytics-vidhya/handling-missing-data-data-pre-processing-8fbab02c8cb4 [Accessed May 13 2022]

Latest Quality, (2017). When to Use Histogram and the Benefits to Your Business. Available from:

https://www.latestquality.com/when-to-use-histogram/ [Accessed, May 12 2022]

Tableau, Guide to Data Cleaning: Definition, Benefits, Components, And How to Clean Your Data. Available from: https://www.tableau.com/learn/articles/what-is-data-cleaning [Accessed May 9 2022]

TDS, (2021). Imputing Numerical Data: Top 5 Techniques Every Data Scientist Must Know. Available from: https://towardsdatascience.com/imputing-numerical-data-top-5-techniques-every-data-scientist-must-know-587c0f51552a [Accessed May 9 2022]

The Information Lab, (2015). Show Me How: Side-by-side circles. Available from:

https://www.theinformationlab.co.uk/2015/03/17/show-side-side-circles [Accessed May 13 2022]