Northumbria University
NEWCASTLE

Newcastle • London • Amsterdam

COURSEWORK COVER PAGE

Module Number:	LD7186
Module Title:	Big Data Analytics
Module Tutor Name: ** delete as appropriate*	SYED RAZA [SR]
Coursework Title:	Assignment
Student Name:	Rasikh Sadiq Thakur
Student ID:	W23042021
Programme of Study:	MSc Big Data and Data Science Technologies with Advanced Practice [BD]
Word count:	5117

Submission Instructions

1. Name your submission in according to the name convention, LD7186_<your tutor initial>_<your programme><sem>_<your id><your first name>.docx, eg LD7186_DC_CT2_w22012345John.docx is the filename for the student enrolled in MSc Big Data & Data Science Tech Sem 2 (2023 Sep Intakes) attending Dileks's session.
2. Submit to Final Report Submission Point at Bb before 16:00, 28 May 2024

Declaration

I confirm that this assessment is my own work and that I have duly acknowledged and correctly referenced the work of others. I am aware of and understand that any breaches to the Code of Academic Conduct will be investigated and sanctioned in accordance with the Academic Conduct Regulation.

Your signature:	Rasikh Sadiq Thakur	**Date:**	23/05/2024

Contents

Section 1 - Big Data Analytics (Python)

Task 1: Problem Domain, Data Description, and Research Questions

Literature Review:

Introduction:

The use of big data analytics has, therefore, brought about a dramatic change in the aviation industry. To optimize customer satisfaction and gain efficiency in their business procedures and key decisions, carriers increasingly rely on analysis results. In this current review, the near-advanced application of data analytics in the aviation domain will be outlined along with an elaboration of the dataset used for the study.

State-of-the-Art Applications in the Aviation Industry

1. **Customer Experience Enhancements:**

   Optimizing customer experience is perhaps the most common application of big data analytics in the aviation business. Airline companies study a tonne of data concerning consumers to find out about their preferences, issues, and their benchmarks of satisfaction. For instance, Li et al. (2017) demonstrated how sentiment analysis of the social media data could reveal the passengers' sentiments and help the airlines to prevent issues. Similarly, the feedback information of consumers is employed as a way of promoting the use of certain products, developing individual solutions for clients, and increasing overall satisfaction. In 2019. Nguyen et al.

2. **Operational Efficiency:**

   The use of data analytics is relevant because the aviation industry involves the coordination of several aspects that require analysis. To minimize the overall working time and, consequently, increase flight availability, the airlines, for example, use predictive analytics for making forecasts regarding necessary maintenance (Ayra & Desai, 2017). Thus, the route optimization algorithms help the airlines in the reduction of both fuel cost and operating costs by identifying the most efficient flight connections from among the numerous ones available.

3. **Revenue Management:**

   Evaluations of aviation companies also indicated that they employed data analytical tools significantly in other areas, such as revenue management. In spite of this, arguments have shown that coming up with dynamic prices and adjusting ticket prices according to the predictions of demand maximizes revenue (McGill & Van Ryzin, 1999). Moreover, it's noteworthy that due to the use of the individualized incentives/reward strategy, the data-driven loyalty programs help to maintain the involvement of airlines' high-value customers (Smith et al., 2017).

4. **Market and Competitive Analysis:**

   By means of data analytics airlines can examine their competitors and the market situation. Airlines may resolve extensions of routes, alliances, and marketing strategies with cognizance because of analysis of trends in the industry, consumer preference, and competitive methods of the rivals (Mayer, 2018).

Problem Domain and Data Description:

This data set, "Airline_Review.csv", seems to be dealing with the domain of air passengers, reviews, and comments. Well, it is one of the forms of customer relations with airlines since it can collect users' data on the airline type, flight routes, feedback, reviews, and customer experiences.

Detailed description of the dataset:

The dataset appears to contain the following columns:

Passenger_Name: Name of the passenger.
Flying_month: Month of the flight.
Route: Route of the flight (departure and arrival destinations).
Rating: Rating provided by the passenger (likely on a scale of 1 to 10).
Verified: Verification status of the review.
Review_title: Title of the review.
Review_content: Detailed content of the review.
Traveller_type: Type of the traveller.
Class: Class of travel (e.g. Economy, Business, Premium Economy).

Research Questions:
Based on the dataset, we can formulate the following research questions:

1. What is the overall distribution of customer satisfaction ratings for flights in the dataset?
   Objective: To analyze and visualize the distribution of customer satisfaction ratings across all flights in the dataset.
2. How do different travel routes affect customer satisfaction ratings?
   Objective: To examine whether certain routes have higher or lower average satisfaction ratings and identify any patterns or trends based on the route.
3. What impact do specific service issues have on customer satisfaction?
   Objective: To identify which service issues are most frequently mentioned in negative reviews and assess their impact on overall satisfaction ratings.
4. Are there differences in customer satisfaction based on the travel class?
   Objective: To compare satisfaction ratings across different travel classes to see if passengers in higher classes report better experiences.

Hypothesis:
Hypothesis: Passengers traveling in Business Class will have significantly higher satisfaction ratings compared to those traveling in Economy Class.

Null and Alternate Hypothesis:
Null Hypothesis(H0): There is no significant difference in customer satisfaction ratings between passengers traveling in Business Class and those traveling in Economy Class.

Alternate Hypothesis(H1): Passengers traveling in Business Class have significantly higher satisfaction ratings compared to those traveling in Economy Class.

Formal Representation:
H0: $\mu_{BusinessClass} = \mu_{EconomyClass}$
H1: $\mu_{BusinessClass} > \mu_{EconomyClass}$

Where $\mu_{BusinessClass}$ represents the mean satisfaction rating for passengers in Business Class, and $\mu_{EconomyClass}$ represents the mean satisfaction rating for passengers in Economy Class.

Task 2: Solution Exploration

Evaluation of Approaches and Technologies:

1. **Python with pandas, Numpy, and Matplotlib/Seaborn:**
   Python is a popular choice because of its adaptability and strong packages designed especially for data processing tasks:

   Pandas: It is an important tool for cleaning, and manipulating, and where some sort of exploratory data analysis (EDA) might be performed. It enhances the easy use of functions and data structures when it comes to stream sorting and manipulation of structured data.

NumPy: Furnishes first-class mathematical capabilities and array computations that can enhance a plethora of calculations. It will be useful for handling large volumes of data and performing the calculations to be done on the data to arrive at an analysis.

Matplotlib/Seaborn: Seaborn can be considered as an extension to this Python package for data visualization called Matplotlib. Matplotlib has a wider variety of plots and beneath the options of Seaborn, it offers more opportunities to adjust them.

2. **Big Data Frameworks like Apache Spark:**

   Leading big data framework Apache Spark is well-known for its scalability and speed, making it ideal for processing massive datasets across distributed computer clusters:

   Distributed Processing: Spark processes big data volumes better than the traditional processing frameworks since it can chop up the tasks of data processing into sub-problems and distribute them among computers that are in a cluster.

   Scalability: Ideal for applications that require processing data that is in petabytes because it is designed to scale up from a single server to thousands of servers.

3. **Machine Learning Algorithms:**

   Algorithms for machine learning (ML) are essential for predictive analytics and data extraction:

   Logistic Regression: Logistic regression is a simple method used in binary classifiers; it involves using past data to forecast client attrition or the probability of a purchase.

   Random Forests: Random forests can address complex data samples and enhance the entropy values for improved predicted accuracy is useful for both classification and regression tasks.

Solutions and Techniques for Similar Problems:

Diverse approaches have been employed to tackle comparable issues within the realm of consumer feedback research:

1. **Sentiment Analysis:**

   Sentiment analysis is the practice of identifying the sentiment represented in textual data, such as customer reviews or social media comments, by applying natural language processing (NLP) techniques:

   Application: From the analysis of consumers’ perceptions, attitudes, and sentiments, airlines may be in a position to identify areas, s of service deliver and the general customer experience or satisfaction.

2. **Classification Algorithms:**

   Data points are categorized using classification algorithms into predetermined classes according to characteristics and past trends:

   Logistic Regression: Using previous reservation data, the airlines are likely to forecast demand along a particular route or the likelihood that a flight would be full.

   Random Forests: These are applied in customer-oriented strategies or service promotions in attempting to predict customer needs or actions from related demographics and past interactions.

3. **Topic Modeling:**

   Latent Dirichlet Allocation (LDA) is one topic modeling approach that is used to find underlying topics or themes in vast amounts of textual data:

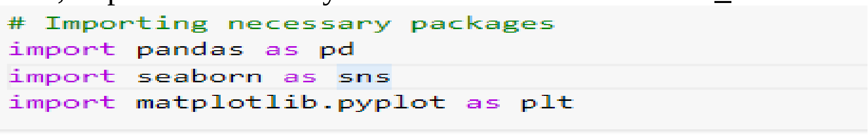
Tool: Since LDA can help to see the tendencies of the customers' feedback, for example, having many complaints heard or several positive aspects of the provided service, airlines will be able to manage the changes for the better.

Chosen Methodological Approach:
Python for Comprehensive Analysis:
Python will be our main tool along with the Pandas, NumPy and Matplotlib/Seaborn libraries do some data processing. This strategy has many benefits:

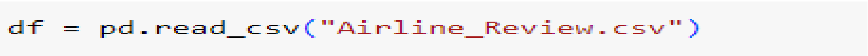
1. **Flexibility:** Python has this meta property since it is flexible and can be used in the areas of data analytics, visualization, and manipulation, thereby customizing our study to the research questions and dataset.
2. **Rich Ecosystem:** The existence of a large number of Python modules and tools for numerous data analytics tasks affirms that the ecosystem has everything we need to solve the problem.
3. **Ease of Interpretation:** The Python in the teamwork as well as the present results communication is OK, it is less difficult to read and understand.

Justification:
Python's strong data processing, analysis, and visualization capabilities make it the chosen technology for this study:
Support: Due to the availability of extensive toolkits in the context of Python and its libraries, data exploration, hypothesis testing, and data visualization also become rigid to have a better understanding of the results and findings of the study on a specific set of data.

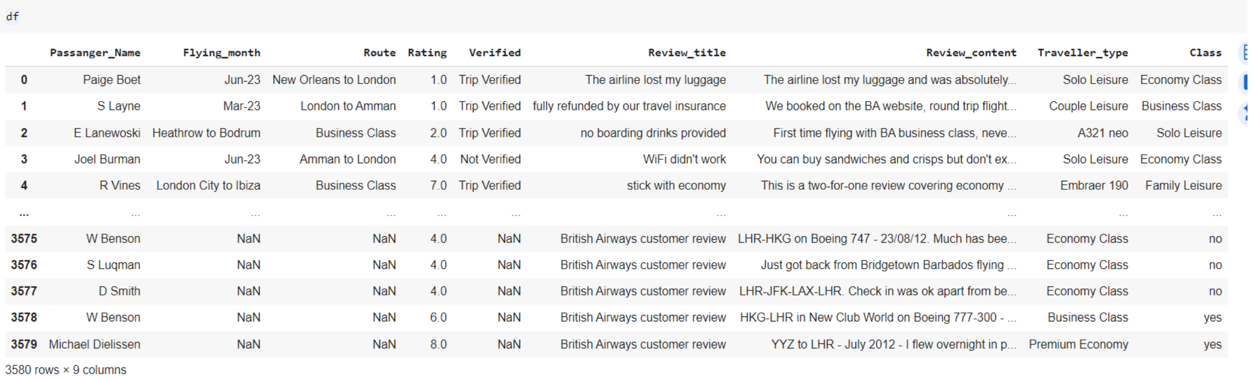
Task 3: Solution Development
First, import the necessary libraries and read the Airline_Review.csv data.



```
# Importing necessary packages
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```



```
df = pd.read_csv("Airline_Review.csv")
```

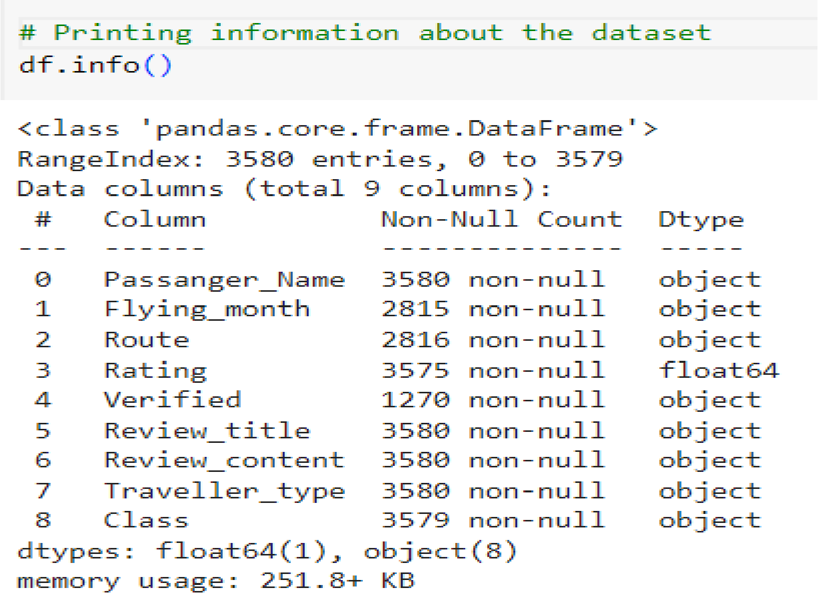


```
df
```

	Passanger_Name	Flying_month	Route	Rating	Verified	Review_title	Review_content	Traveller_type	Class
0	Paige Boet	Jun-23	New Orleans to London	1.0	Trip Verified	The airline lost my luggage	The airline lost my luggage and was absolutely...	Solo Leisure	Economy Class
1	S Layne	Mar-23	London to Amman	1.0	Trip Verified	fully refunded by our travel insurance	We booked on the BA website, round trip flight...	Couple Leisure	Business Class
2	E Lanewoski	Heathrow to Bodrum	Business Class	2.0	Trip Verified	no boarding drinks provided	First time flying with BA business class, neve...	A321 neo	Solo Leisure
3	Joel Burman	Jun-23	Amman to London	4.0	Not Verified	WiFi didn't work	You can buy sandwiches and crisps but don't ex...	Solo Leisure	Economy Class
4	R Vines	London City to Ibiza	Business Class	7.0	Trip Verified	stick with economy	This is a two-for-one review covering economy ...	Embraer 190	Family Leisure
...	...	...	...	...	...	...	...	...	...
3575	W Benson	NaN	NaN	4.0	NaN	British Airways customer review	LHR-HKG on Boeing 747 - 23/08/12. Much has bee...	Economy Class	no
3576	S Luqman	NaN	NaN	4.0	NaN	British Airways customer review	Just got back from Bridgetown Barbados flying ...	Economy Class	no
3577	D Smith	NaN	NaN	4.0	NaN	British Airways customer review	LHR-JFK-LAX-LHR. Check in was ok apart from be...	Economy Class	no
3578	W Benson	NaN	NaN	6.0	NaN	British Airways customer review	HKG-LHR in New Club World on Boeing 777-300 - ...	Business Class	yes
3579	Michael Dielissen	NaN	NaN	8.0	NaN	British Airways customer review	YYZ to LHR - July 2012 - I flew overnight in p...	Premium Economy	yes

3580 rows × 9 columns

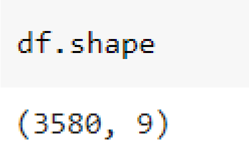
After this have printed the information about the dataset using the info() function.



```
# Printing information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3580 entries, 0 to 3579
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Passanger_Name  3580 non-null   object 
 1   Flying_month    2815 non-null   object 
 2   Route           2816 non-null   object 
 3   Rating          3575 non-null   float64
 4   Verified        1270 non-null   object 
 5   Review_title    3580 non-null   object 
 6   Review_content  3580 non-null   object 
 7   Traveller_type  3580 non-null   object 
 8   Class           3579 non-null   object 
dtypes: float64(1), object(8)
memory usage: 251.8+ KB
```

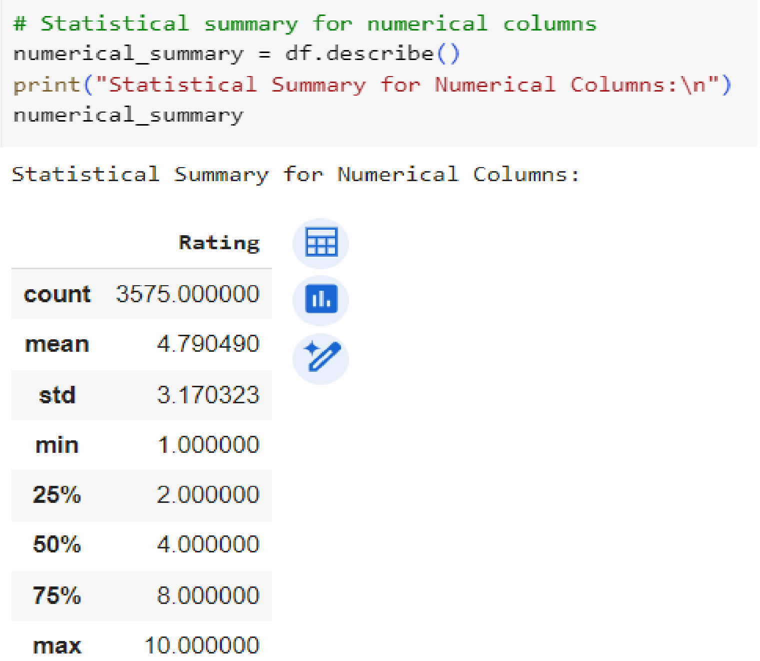
Printed the shape of the dataset which we got as 3580 rows and 9 columns.



```
df.shape
```

```
(3580, 9)
```

After this, we have the Statistical Summary of the Numerical column

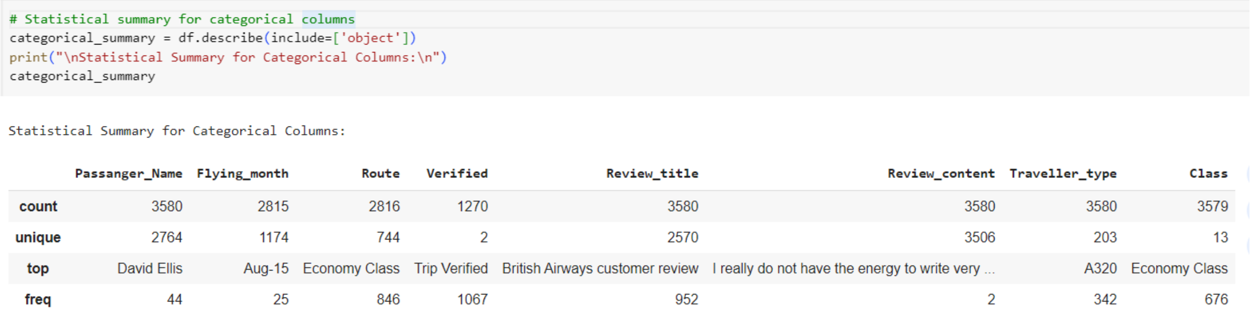


```
# Statistical summary for numerical columns
numerical_summary = df.describe()
print("Statistical Summary for Numerical Columns:\n")
numerical_summary
```

```
Statistical Summary for Numerical Columns:
```

	Rating
count	3575.000000
mean	4.790490
std	3.170323
min	1.000000
25%	2.000000
50%	4.000000
75%	8.000000
max	10.000000

And then a statistical summary for the categorical column.



```
# Statistical summary for categorical columns
categorical_summary = df.describe(include=['object'])
print("\nStatistical Summary for Categorical Columns:\n")
categorical_summary
```

```
Statistical Summary for Categorical Columns:
```

	Passanger_Name	Flying_month	Route	Verified	Review_title	Review_content	Traveller_type	Class
count	3580	2815	2816	1270	3580	3580	3580	3579
unique	2764	1174	744	2	2570	3506	203	13
top	David Ellis	Aug-15	Economy Class	Trip Verified	British Airways customer review	I really do not have the energy to write very ...	A320	Economy Class
freq	44	25	846	1067	952	2	342	676

Checking for null values in the dataset

```
# Checking if there are any null values in the dataset
df.isnull().sum()
```

```
Passanger_Name        0
Flying_month        765
Route               764
Rating                5
Verified           2310
Review_title          0
Review_content        0
Traveller_type        0
Class                 1
dtype: int64
```

Handling null/missing values in the dataset

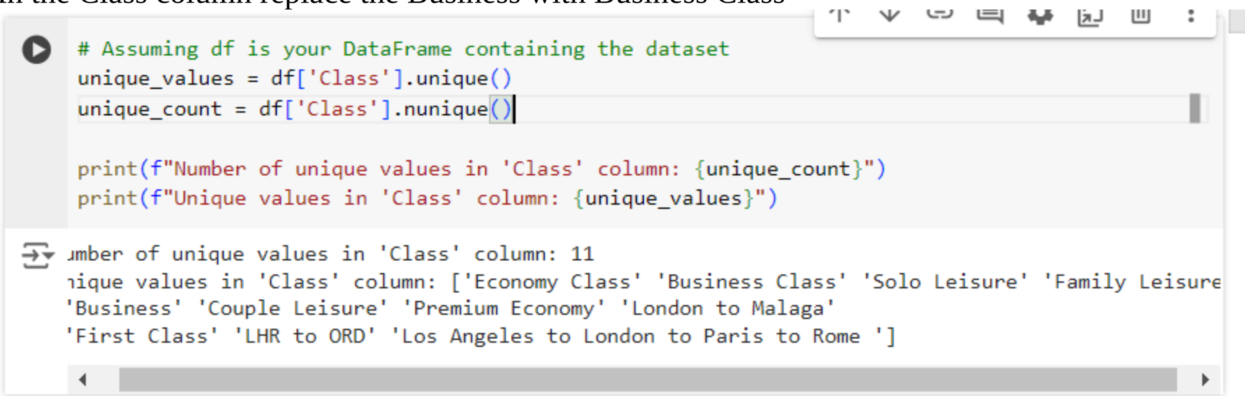
```
# Drop rows with missing values in 'Flying_month' and 'Route' columns
df.dropna(subset=['Flying_month', 'Route'], inplace=True)
```

```
# Impute missing values in 'Rating' column with median
median_rating = df['Rating'].median()
df['Rating'].fillna(median_rating, inplace=True)
```

```
# Impute missing value in 'Class' column with most frequent class
most_frequent_class = df['Class'].mode()[0]
df['Class'].fillna(most_frequent_class, inplace=True)
```

```
# Replace missing values in 'Verified' column with 'Unknown'
df['Verified'].fillna('Unknown', inplace=True)
```

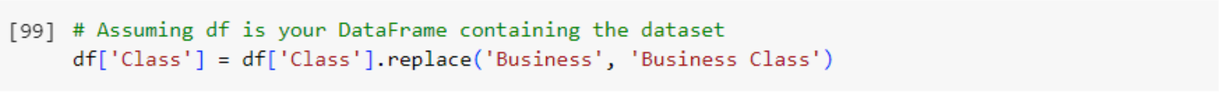
In the Class column replace the Business with Business Class



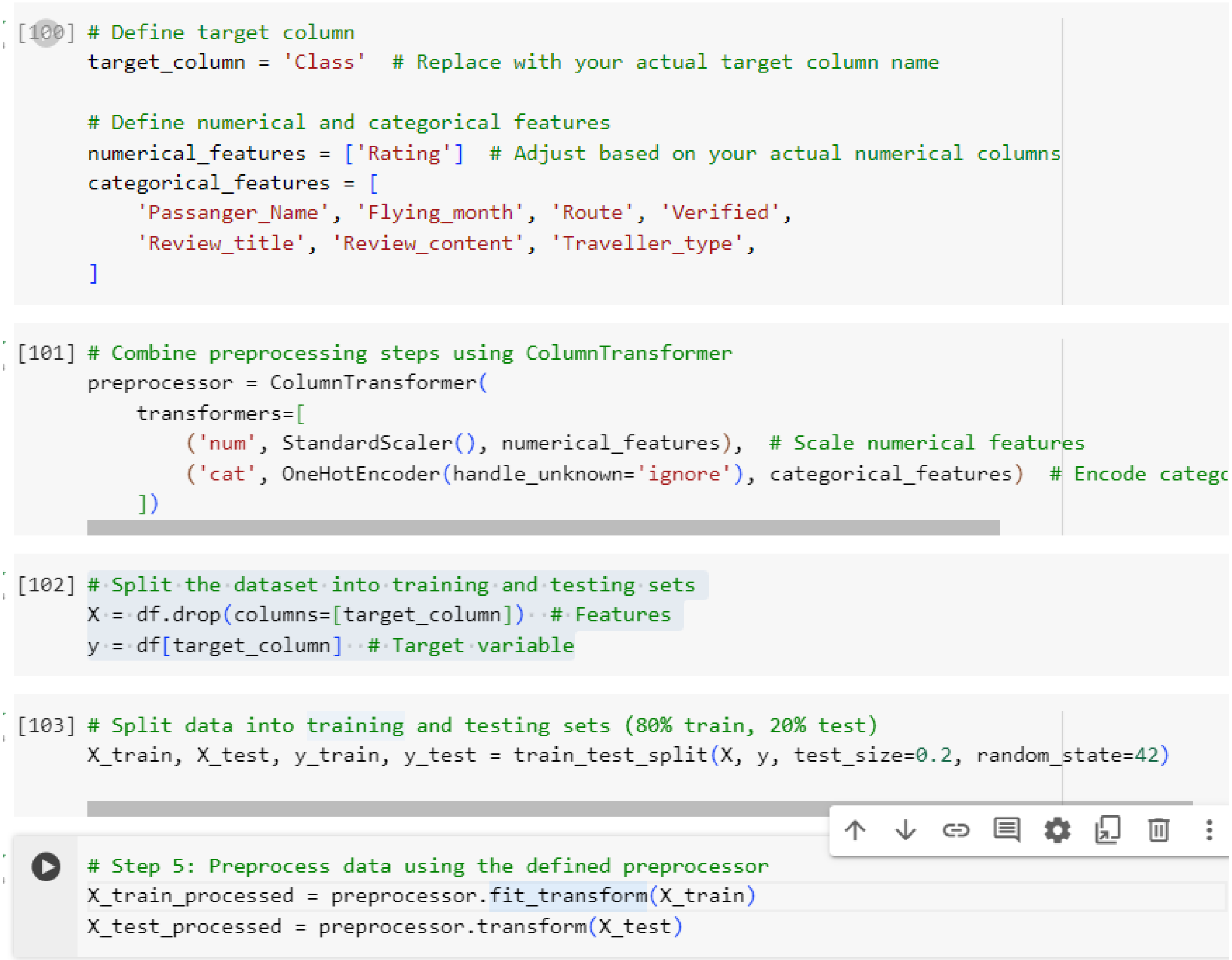
```
# Assuming df is your DataFrame containing the dataset
unique_values = df['Class'].unique()
unique_count = df['Class'].nunique()

print(f"Number of unique values in 'Class' column: {unique_count}")
print(f"Unique values in 'Class' column: {unique_values}")
```

```
umber of unique values in 'Class' column: 11
ique values in 'Class' column: ['Economy Class' 'Business Class' 'Solo Leisure' 'Family Leisure
'Business' 'Couple Leisure' 'Premium Economy' 'London to Malaga'
'First Class' 'LHR to ORD' 'Los Angeles to London to Paris to Rome ']
```



```
[99] # Assuming df is your DataFrame containing the dataset
     df['Class'] = df['Class'].replace('Business', 'Business Class')
```



```
[100] # Define target column
      target_column = 'Class'  # Replace with your actual target column name

      # Define numerical and categorical features
      numerical_features = ['Rating']  # Adjust based on your actual numerical columns
      categorical_features = [
          'Passanger_Name', 'Flying_month', 'Route', 'Verified',
          'Review_title', 'Review_content', 'Traveller_type',
      ]
```

```
[101] # Combine preprocessing steps using ColumnTransformer
      preprocessor = ColumnTransformer(
          transformers=[
              ('num', StandardScaler(), numerical_features),  # Scale numerical features
              ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)  # Encode catego
          ])
```

```
[102] # Split the dataset into training and testing sets
      X = df.drop(columns=[target_column])  # Features
      y = df[target_column]  # Target variable
```

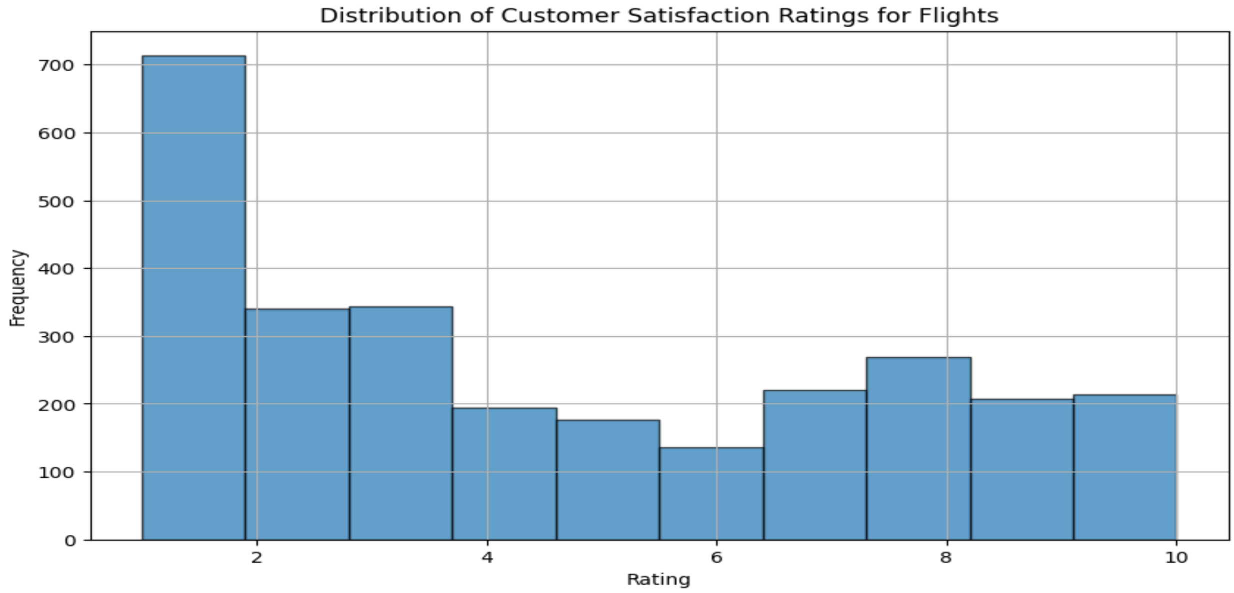
```
[103] # Split data into training and testing sets (80% train, 20% test)
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Step 5: Preprocess data using the defined preprocessor
X_train_processed = preprocessor.fit_transform(X_train)
X_test_processed = preprocessor.transform(X_test)
```

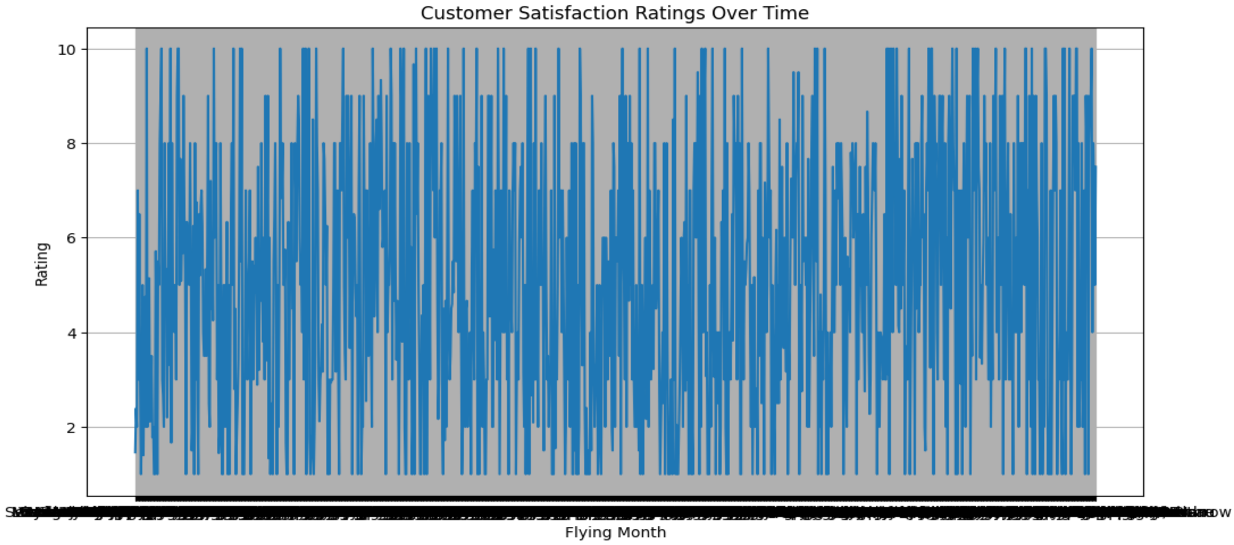
Therefore, data pretreatment is necessary to ensure that the data collected is appropriate and of good quality for machine learning research. To maintain cleanliness, for the first operation, missing values were addressed using methods such as using the median on numeric attributes or eradicating partial rows as required. Next, as with any model whose parameters might contain suspicious nuggets of a specific magnitude, numerical attributes such as "Rating" were normalized by applying the StandardScaler to make their scales uniform.

We employed one-hot encoding for the categorical data namely; 'Passanger_Name" and 'Flying_month' since data mining methods require quantitative rather than qualitative data. In this method, several binary columns are developed for each categorical characteristic, where each column indicates if the given category exists in the observations. Also, for the validity of the evaluations, we split the whole collection of files into training and testing datasets. The prediction accuracy of the model on unseen data was measured with the help of the testing set to avoid bias, while the training set was used to train the model. These preprocessing stages were done in harmony with applying the ColumnTransformer of scikit-learn ensuring a fast and consistent process along the analysis.

After preprocessing the dataset. Now doing Research questions analysis.
Research Question 1:
What is the overall distribution of customer satisfaction ratings for flights in the dataset?

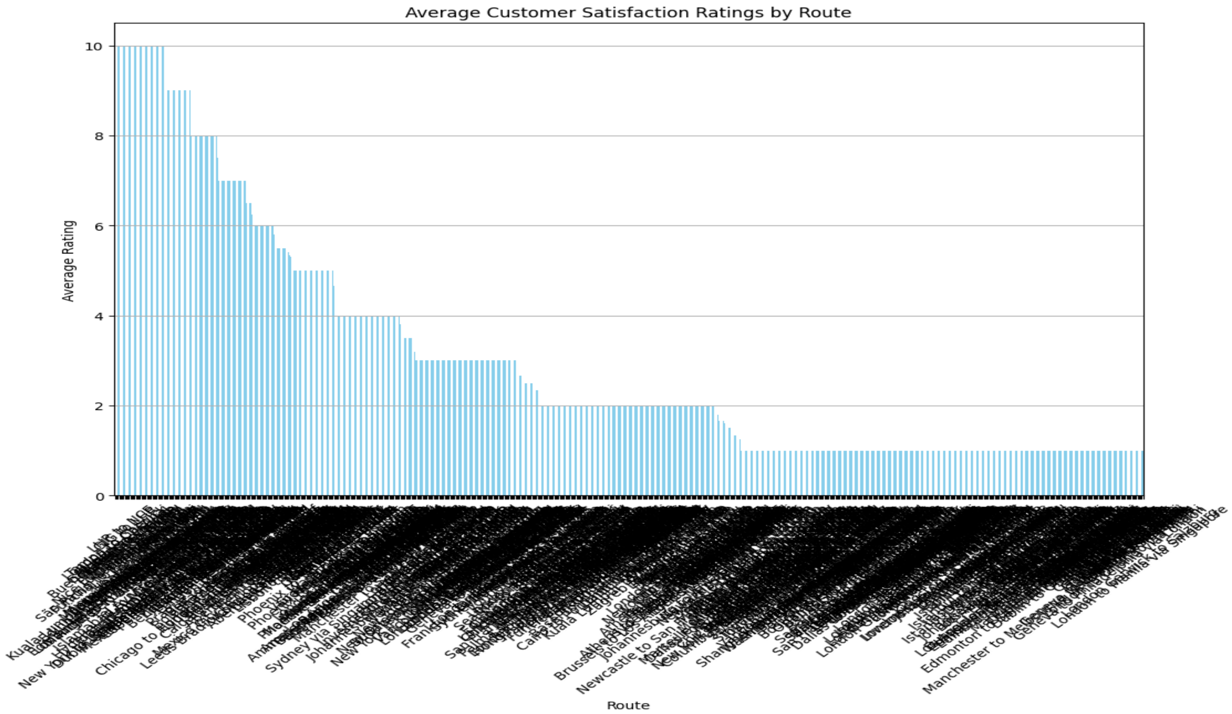


The histogram presents the distribution of the passenger satisfaction ratings concerning the flights ranging from 1 to 10. This reveals that a majority of the customers are over 700, the most frequent rating is 1 meaning customers are highly dissatisfied. It means that the clients with ratings of 2 and 3 have fairly high frequencies, namely 316 and 302 respectively. Looking into the ratings 4 up to 6, there is a larger gap in the occurrence of respondents with mid-level satisfaction evaluation. Consumers with ratings 7 through 10 are more balanced; each of them shared between 200 and 300 replies and can be considered somewhat satisfied. The individual approach also reveals that the augmented distribution of negative feedback contains more substantial bad ratings, and this confirms consumers’ dissatisfaction.



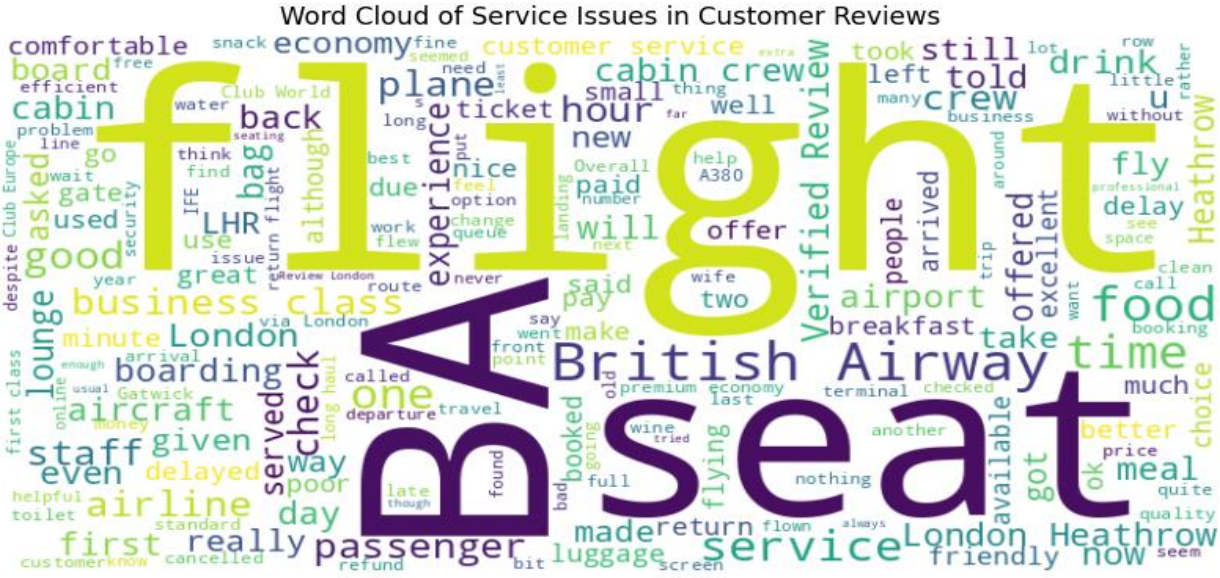
Quantitative results for flyers’ satisfaction ratings are presented in the flying months in a line graph with the scores on the Y-axis starting from 1 to 10. In respect of a given month, the rating distribution is represented by the vertical line. The results show that within all the periods the consumers’ experiences differ significantly, and the density and distribution of the evaluations is also quite dense and heterogeneous in each of the months. Here, the variation of the high and low ratings indicated in the figure does not exhibit any pattern of progression over the studied period. This signifies that there is no direct correlation between the season and the measures of happiness and that the consumer’s happiness is relatively volatile.

Research Question 2:
How do different travel routes affect customer satisfaction ratings?

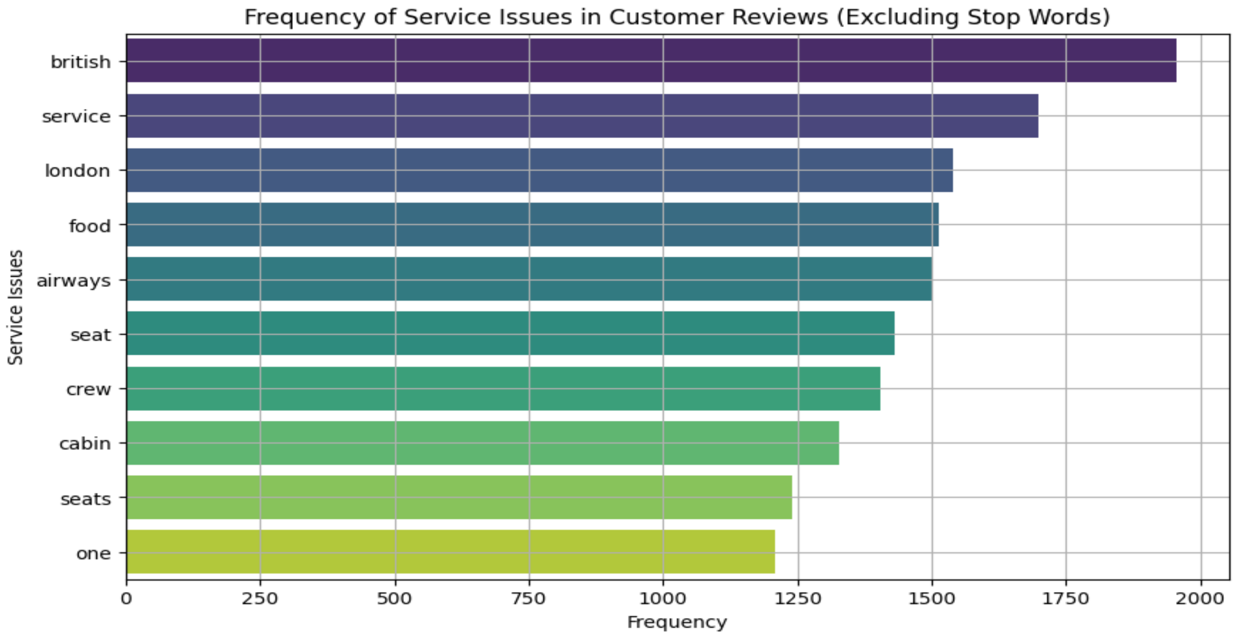


The bar chart provides the average customer satisfaction ratings for the given routes for each bar represents a different route. The routes are mentioned along the x-axis and the ratings are between 0 and 10 along the y-axis. The bars on the other hand depict that, some routes are far better when it comes to level of satisfaction than others and they are aligned in order of average rating. Again the labels on the horizontal axis aren’t very clear because they are written tilted and written very close to each other. The minimum ratings recorded are approximately equal to 0, whereas the maximum ratings are slightly above 10 which shows marked disparities occurring in different routes regarding the satisfaction level of the consumers.

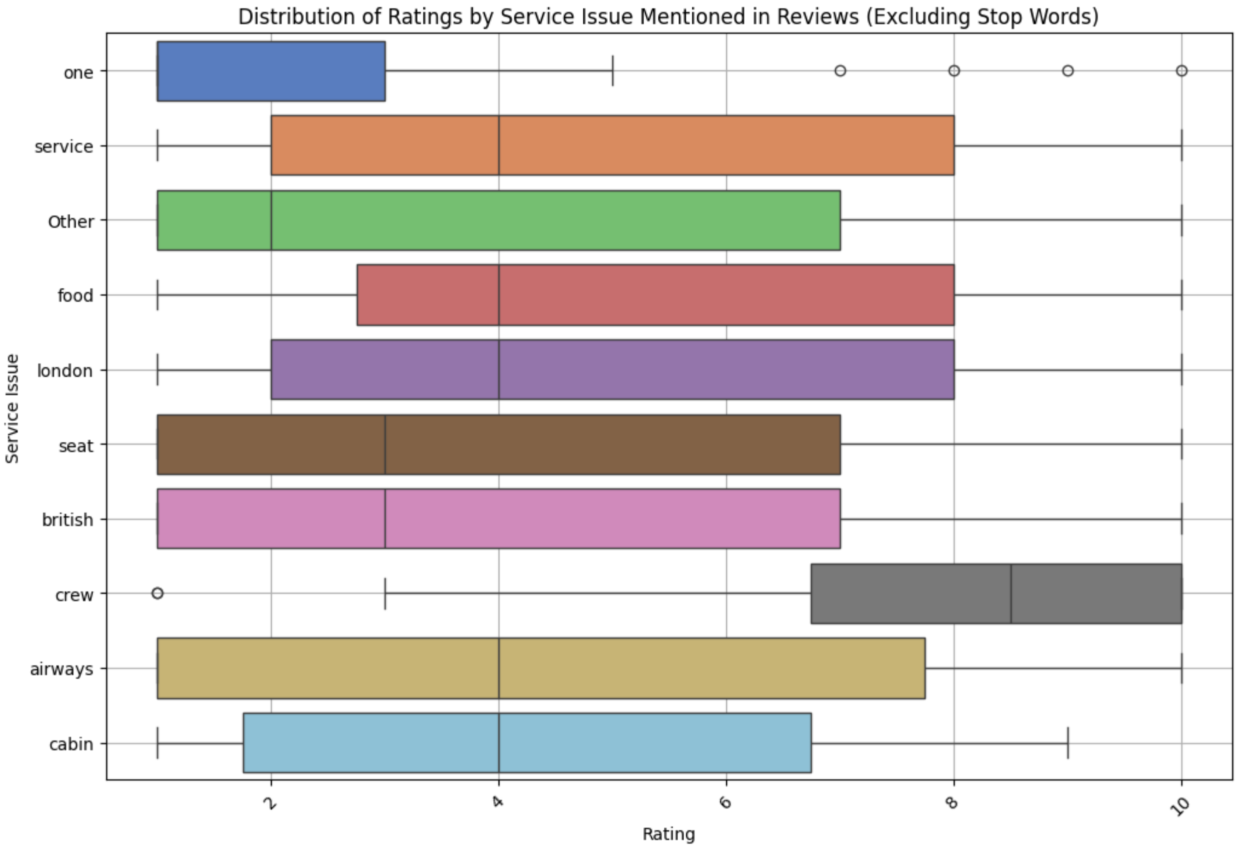
Research Question 3:
What impact do specific service issues have on customer satisfaction?



The word cloud below shows how often terms related to service are used in customers’ reviews; terms like “flight,” “seat,” “BA” (British Airways), “business,” and “class” are emphasized to show areas of interest or concern. The reference to these terms presupposes that issues with business class services, flying experience, and comfortable seats are valuable factors that shape consumers’ satisfaction. The words ‘crew’, ‘service’, ‘delay’, ‘luggage ‘, and ‘food’ also highlight other things that are vital and define the overall consumer experience. That these specific service concerns are mentioned frequently conveys how business-sensitive the different factors influencing passengers' perception and evaluation of their degree of satisfaction with the airline are.

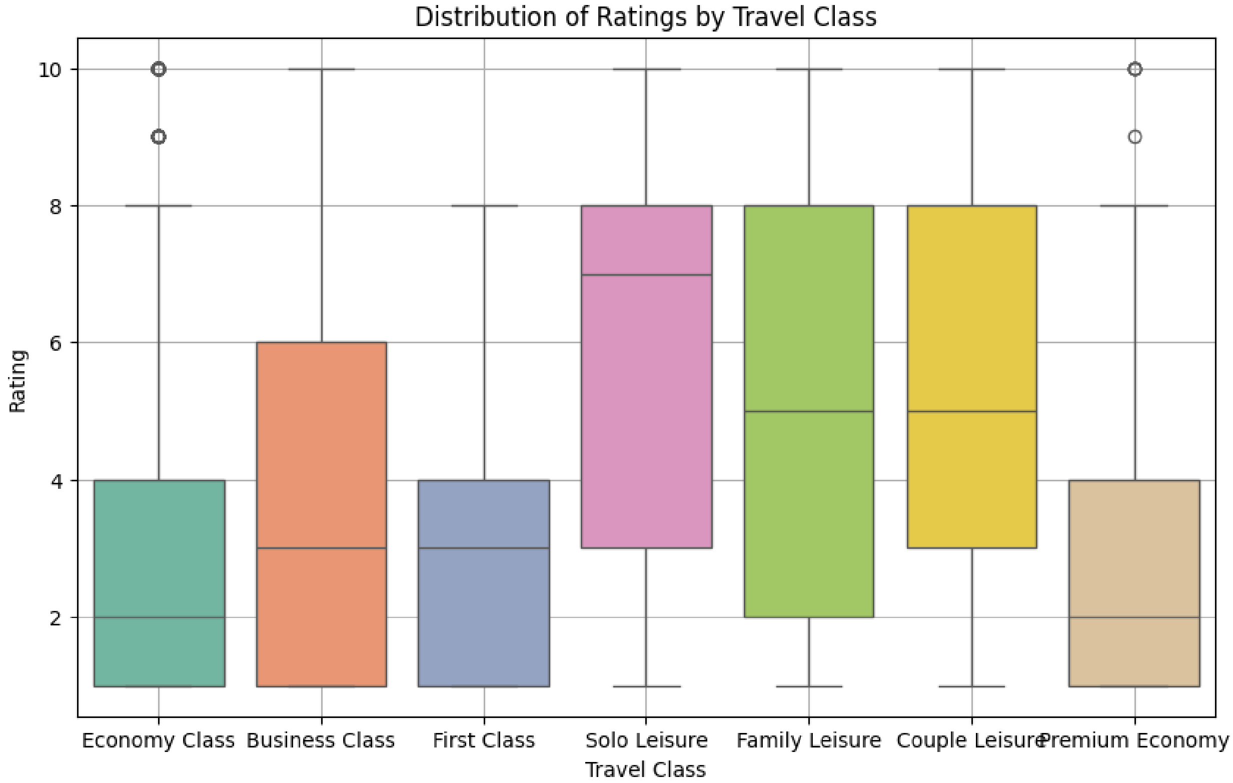


According to the bar chart depicting the frequently stated service concerns in the customer reviews, the most frequently mentioned phrases were, ''British'', ''service'', ''London'', ''food', ''airways'', ''seat'', ''crew'', ''cabin'', ''seats'', and ''one''. This means that concerns about food, layout designs, staff behavior and, overall, service quality significantly affect clients' concerns. This means that these phrases are commonly used implying that these problems affect consumer satisfaction in numerous ways. The efficiency and satisfaction levels of the clients may improve in case the general service issues are addressed.

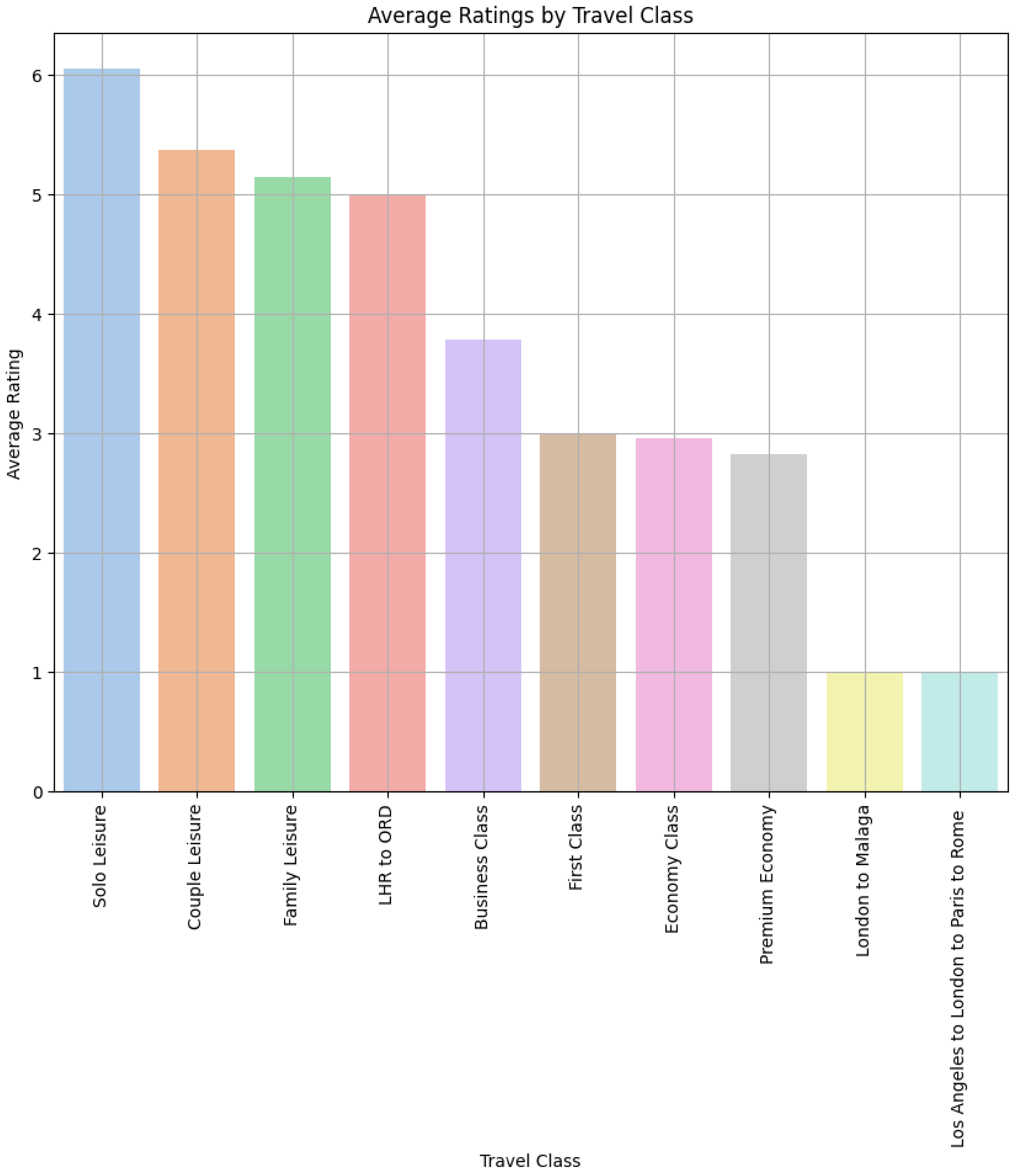


After the stop words are removed, it presents the box plot of the customer ratings concerned with service faults in the reviews. The evaluations reveal a good variation, the crew-related issues receive lower scores thus, suggesting that it spoils customer satisfaction. On the other hand, problems related to a cabin and airways seem to have larger median ratings, which means that it does not influence either positively or negatively as much as the other components. When it comes to sending out and mating calls, extremely unpleasant encounters are illustrated by the 'one' and 'crew' categories. Thus, the graph proves that while some problems are connected with lower satisfaction levels of the services, other problems may not lead to a major deterioration of the experience.

Research Question 4:
Are there differences in customer satisfaction based on the travel class?

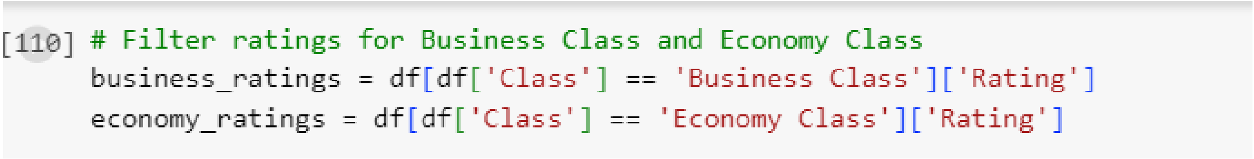


Travel class presents how consumers' satisfaction is influenced as depicted by the boxplot. This is depicted by the poor median ratings registered under the Economy Class and Premium Economy categories. Nevertheless, Business Class and, especially, First Class are way more diverse in terms of the median customer ratings despite being higher. Categories related to leisure travel with the higher median are Couple Leisure, Family Leisure, and particularly Solo Leisure; moreover Solo Leisure seems to reveal higher consistency in the quality experience. In some customer segments such as Economy Class, out of the pattern values, they are infrequent but strongly positive or negative. In general, the improvements in the higher travel classes and dedicated leisure travel segment are associated with a better satisfaction rating from the consumers when compared to the Economy and Premium Economy classes.

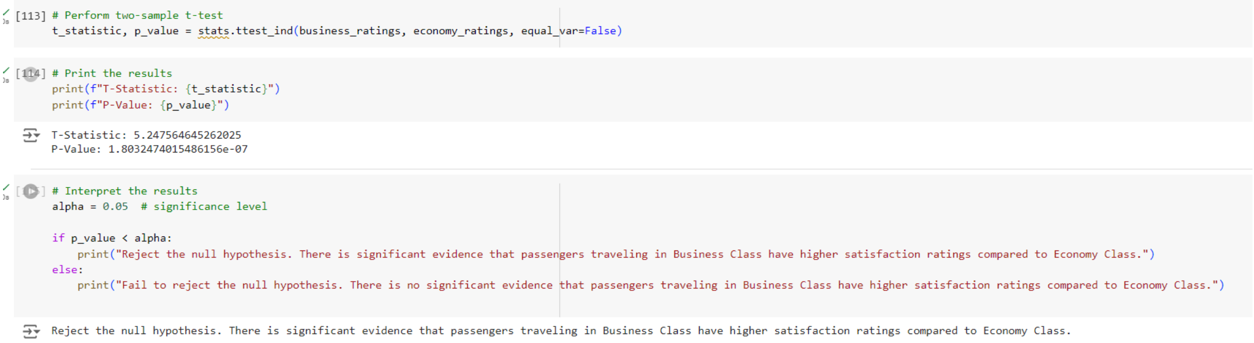


The bar graph suggested that average ratings of leisure traveling categories were higher such as solo travelers, couples, and families having the highest number of happy experiences. Thus, lower satisfaction is depicted by the lowest scores given to the Economy and Premium Economy classes. The satisfaction level of first-class and business-class is moderate. The real-life examples of ‘’LHR to ORD’’ and ‘’Los Angeles to London to Paris to Rome’’ clearly depict that the happiness level of the customer may vary with the travel class as well as the route. As a rule, satisfaction is higher during leisure, and it is lower in Economy and Premium Economy classes.

Hypothesis Explanation



```
[110] # Filter ratings for Business Class and Economy Class
      business_ratings = df[df['Class'] == 'Business Class']['Rating']
      economy_ratings = df[df['Class'] == 'Economy Class']['Rating']
```



```
[113] # Perform two-sample t-test
      t_statistic, p_value = stats.ttest_ind(business_ratings, economy_ratings, equal_var=False)

[114] # Print the results
      print(f"T-Statistic: {t_statistic}")
      print(f"P-Value: {p_value}")

      T-Statistic: 5.247564645262025
      P-Value: 1.8032474015486156e-07

[ ] # Interpret the results
      alpha = 0.05  # significance level

      if p_value < alpha:
          print("Reject the null hypothesis. There is significant evidence that passengers traveling in Business Class have higher satisfaction ratings compared to Economy Class.")
      else:
          print("Fail to reject the null hypothesis. There is no significant evidence that passengers traveling in Business Class have higher satisfaction ratings compared to Economy Class.")

      Reject the null hypothesis. There is significant evidence that passengers traveling in Business Class have higher satisfaction ratings compared to Economy Class.
```

The findings indicate that there is a statistically significant difference with regards to mean customer satisfaction score between Business Class passengers and Economy Class passengers for which hypothesis tests were performed. The p-value is incredibly low and estimated around 1. 8e-07 it specifies a T-statistic of 5.25 hinting at strong evidence supporting one of the hypotheses, the null hypothesis in this case. Therefore, we are in a position to disregard the null hypothesis that stated that there was an insignificant change in performing satisfaction ratings based on the travel class. From this, it can be deduced that travelers in a Business Class are more satisfied than those in the Economy class. Implying the concept that the services need to be differentiated across the travel classes to affect the overall consumer satisfaction in the airline segment.

Task 4: Result Evaluation and Future Development

Result Evaluation:

The following are major findings that were made known when interpreting customer satisfaction ratings in the airline industry. To start with,ũ a definitive distribution prejudice toward lower ratings réalised, which indicates that several passengers are, in general, unsatisfied with certain aspects of their flight experience. The following is another factor: routes; in most instances, specific routes are viewed more favorably in terms of satisfaction. Respondents' attitudes were found to be significantly affected by the quality of the work done by the crew and the comfort of the seats which were perceived to have a direct bearing on passengers' level of satisfaction. Moreover, it was established that the segregation between the classes of travel made a significant impact on the satisfaction levels, where business class travelers recorded far much higher satisfaction than economy class travelers.

Likewise, hypothesis testing affirmed that there is a meaningful difference in passenger satisfaction concerning the Business and Economy Class. This underlines the importance of the need for airlines to adapt their facilities and/or services to different travel classes aspect in a bid to enhance passengers' satisfaction.

Limitations of the Analysis:

1. **Data Quality and Completeness:** The authenticity and totality of the data set become urgently important prerequisites for the study. Bad or deficient data that we use in the analysis might endanger the results' believed accuracy.
2. **Sample Bias:** The fact that the dataset may not accurately capture everybody who traveled by airline plane personally is also likely. Concerns about the biases of reviews' selection or reporting might skew the results and thus make the findings quite narrow.

3. **Causality vs. Correlation:** While the following can lead researchers to correlations among various features, proving cause or effect cannot be done by the use of research. The consumer's preferences and behaviors are further affected by variables that are beyond the dataset like social media influences or personal experiences.

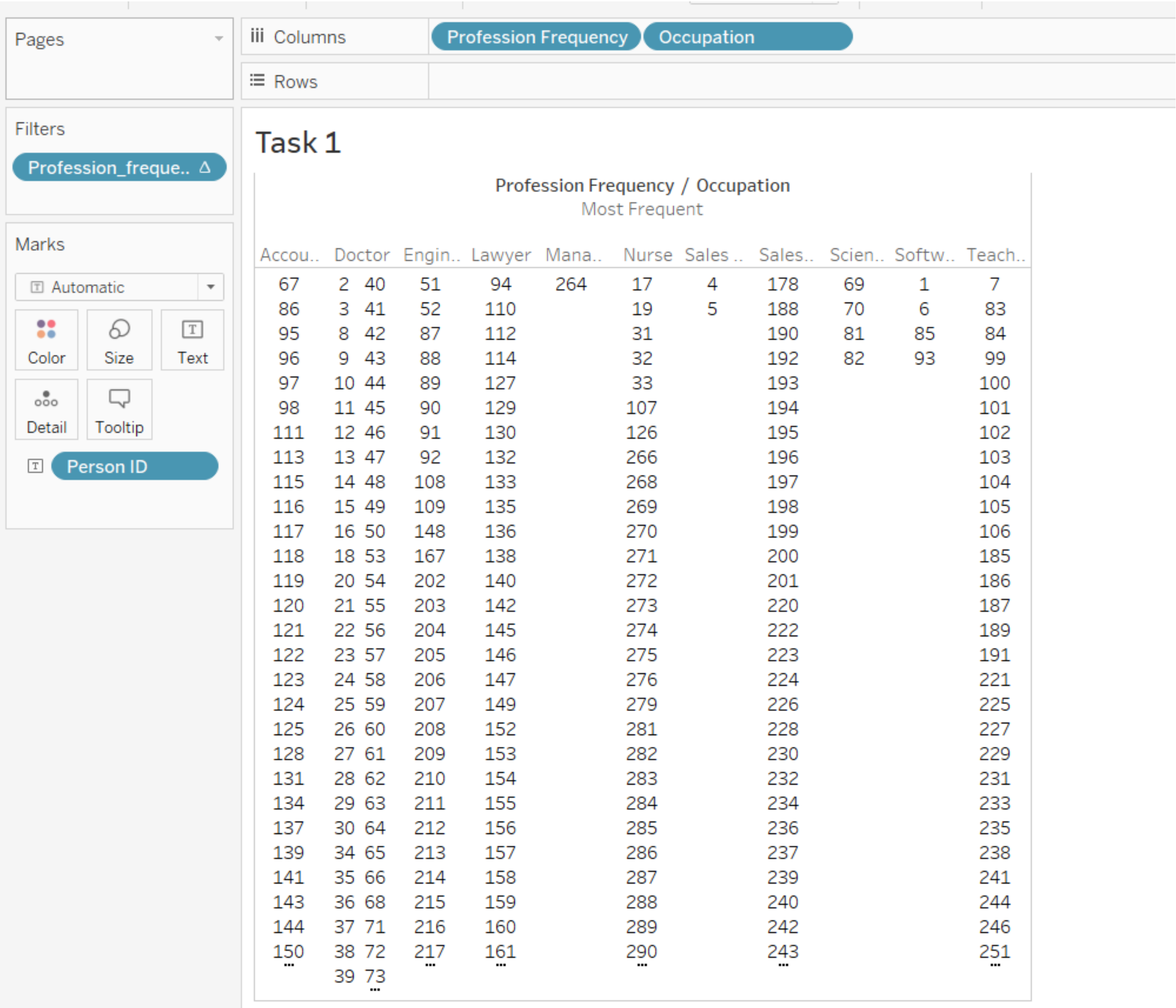
Future Work:

1. **Sentiment Analysis:** Emphasis on knowing the attitudes and feeling of the consumers as they try airline services can be achieved through the implementational of natural language processing and sentiment analysis on their actual descriptions of the services.
2. **Predictive Modeling:** Eventually, the operations of low-cost airlines can be based on the predictive models that anticipate the actions of the customers, i.e. changing their travel selections or canceling their reservations.
3. **Real-Time Data Analytics:** Airlines possibly will catch up and learn how to analyze and interpret information almost instantaneously by employing machine learning and AI features. It would permit to make up the for any gaps and adjust the services, when and how needed.
4. **Integration with Customer Relationship Management (CRM) Systems:** The profiling results, if combined in a way with CRM systems could help in tailor-making marketing campaigns that are based on individual preferences and behavior and also create engagement with customers in a personalized way.

Eventually, the solution will hopefully be able to add extra profitability to the aviation industry by being more successful in terms of customer satisfaction, operational efficiency, and general business success, taking all these factors into account as well as the introduction of innovations in technology and analytics.

Section 2 - Business Intelligence (Tableau)

Task 1



Pages

Columns: Profession Frequency, Occupation

Rows

Filters: Profession_freque.. Δ

Marks: Automatic

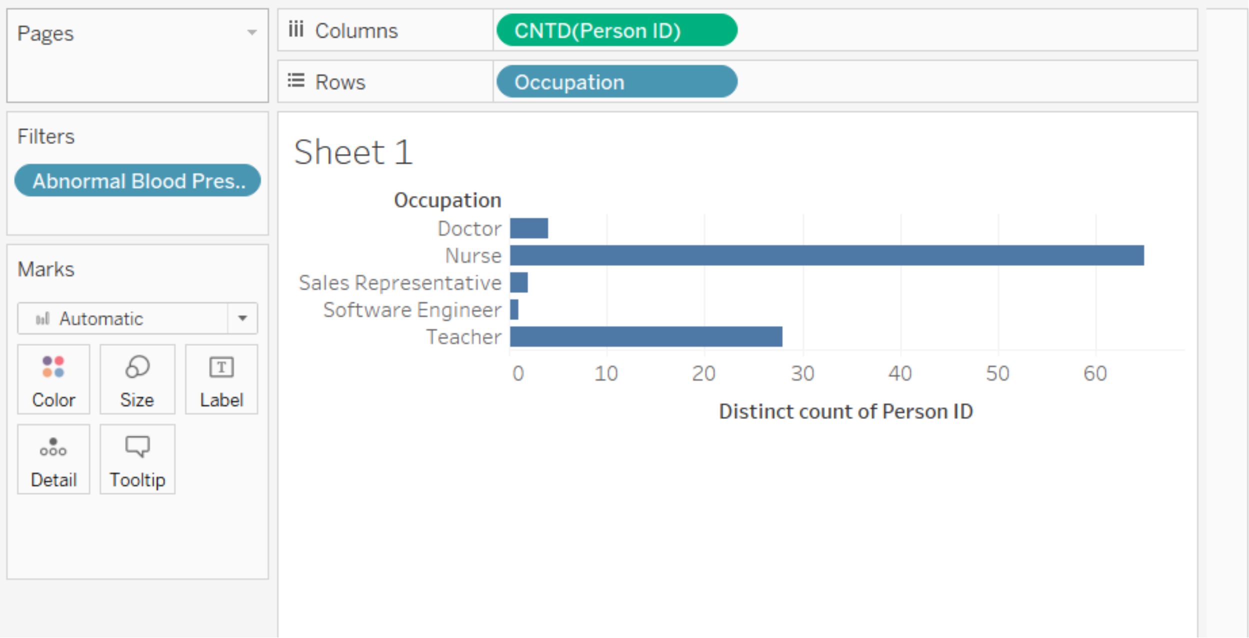
Person ID

Task 1

Profession Frequency / Occupation

Most Frequent

Accou..	Doctor		Engin..	Lawyer	Mana..	Nurse	Sales ..	Sales..	Scien..	Softw..	Teach..
67	2	40	51	94	264	17	4	178	69	1	7
86	3	41	52	110		19	5	188	70	6	83
95	8	42	87	112		31		190	81	85	84
96	9	43	88	114		32		192	82	93	99
97	10	44	89	127		33		193			100
98	11	45	90	129		107		194			101
111	12	46	91	130		126		195			102
113	13	47	92	132		266		196			103
115	14	48	108	133		268		197			104
116	15	49	109	135		269		198			105
117	16	50	148	136		270		199			106
118	18	53	167	138		271		200			185
119	20	54	202	140		272		201			186
120	21	55	203	142		273		220			187
121	22	56	204	145		274		222			189
122	23	57	205	146		275		223			191
123	24	58	206	147		276		224			221
124	25	59	207	149		279		226			225
125	26	60	208	152		281		228			227
128	27	61	209	153		282		230			229
131	28	62	210	154		283		232			231
134	29	63	211	155		284		234			233
137	30	64	212	156		285		236			235
139	34	65	213	157		286		237			238
141	35	66	214	158		287		239			241
143	36	68	215	159		288		240			244
144	37	71	216	160		289		242			246
150	38	72	217	161		290		243			251
…	39	73	…	…		…		…			…
		…									



Pages

Columns: CNTD(Person ID)

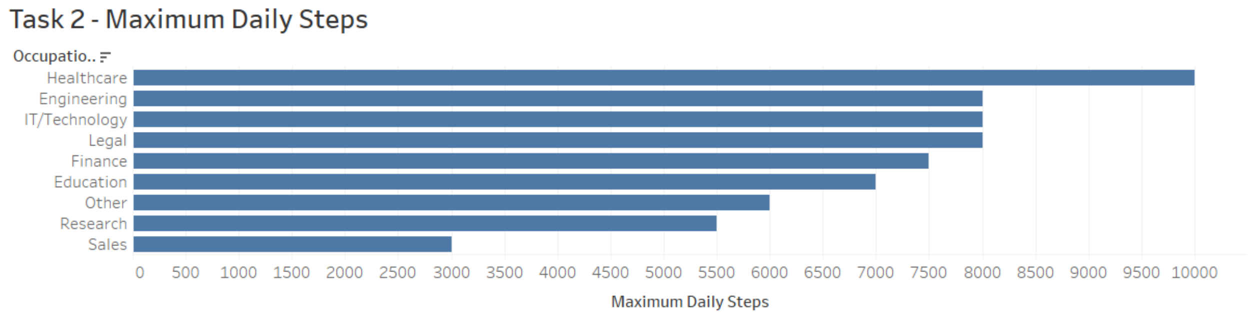
Rows: Occupation

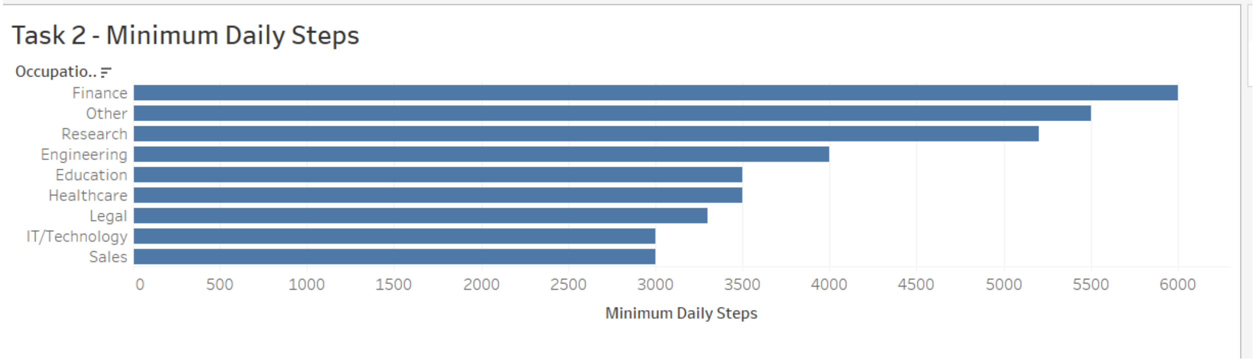
Filters: Abnormal Blood Pres..

Marks: Automatic

The data presented in the blood pressure visualization filters data and spotlights atypical blood pressure in occupations using the 'Profession Frequency' parameter. The "Profession Frequency" field limits the records that are abnormally high in blood pressure, while the "Person ID" field merely shows the number of entries that are available for the group of occupations. The users are allowed to access degree of abnormality of the different professions with hypertension issue either the most severely afflicted or the least through the dynamic displaying of the occupations using Tableau. The understandable interface, sure enough, covers the detailed representation of work-related regular blood pressure variations among the listed positions. The parameter capability of the Tableau can be made more flexible. This can so happen so that the people who are stakeholder may know how the job and the blood pressure are related.

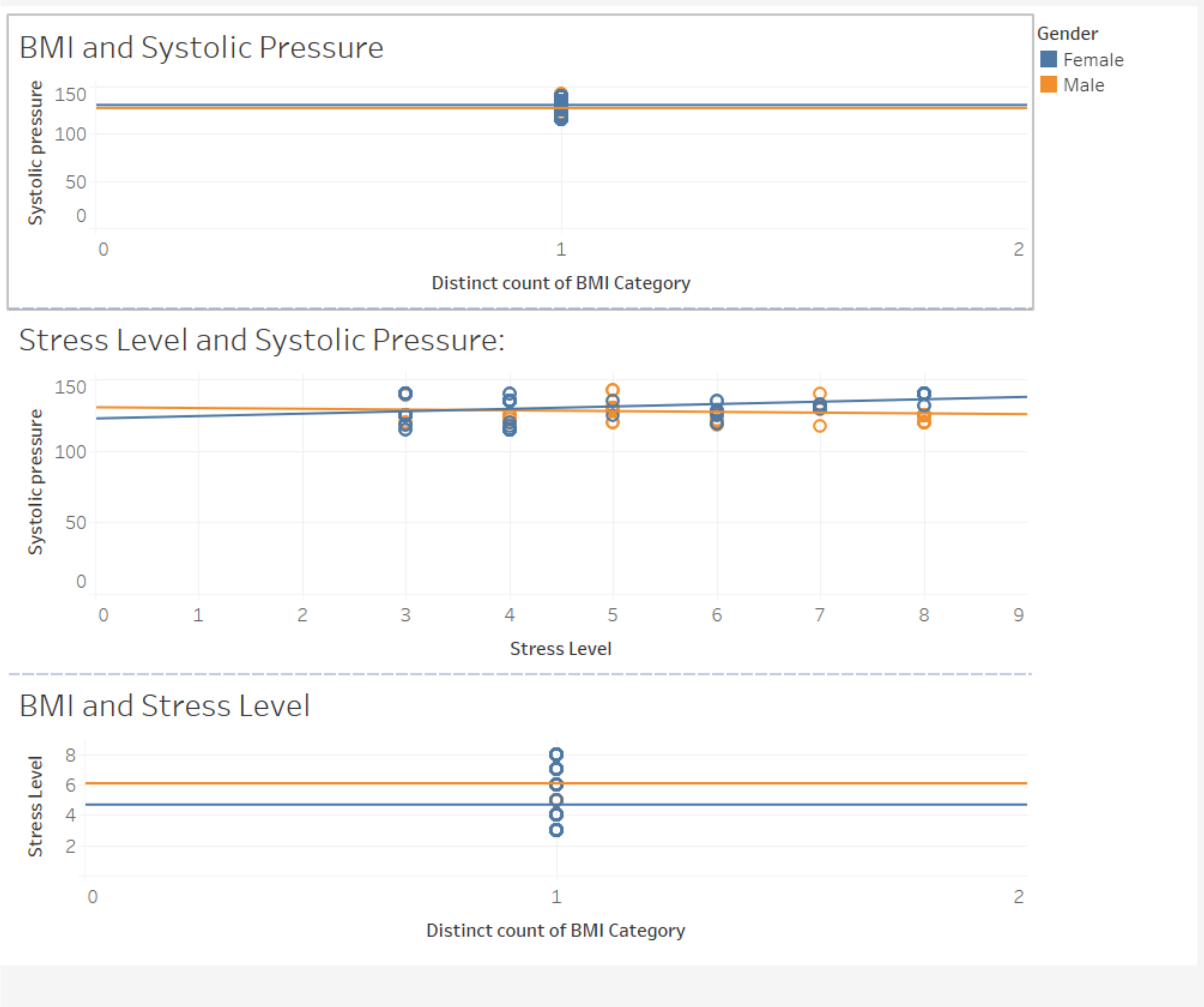
Task 2:





With a combination of a morning rush, a midday lull, and an evening bump, some workers appeared to take more steps during certain hours while most of them had their routines. The bar chart indicating consecutive steps per job type with the biggest and minimal steps - in descending order, was the visualization shown. To create a more general grouping and for the analysis of each category, the "Occupation Category" field was created for roles such as IT/Technology, Healthcare, Sales, and Education. To demonstrate the range between the highest and the lowest step counts the originally computed fields such as “MADAYSTEP” and “MINDAYSTEP” were listed in columns and were reallocated in descending order. The ensuing bar chart, where bars stand for the types of work and their widths represent the variation in the number of steps taken per day, is a comparatively simple graphic to understand. The pictorial representation of the ranges daily steps by day highlights the readability of the visualization and gives a quick grasp opportunity of the data by the public.

Task 3:

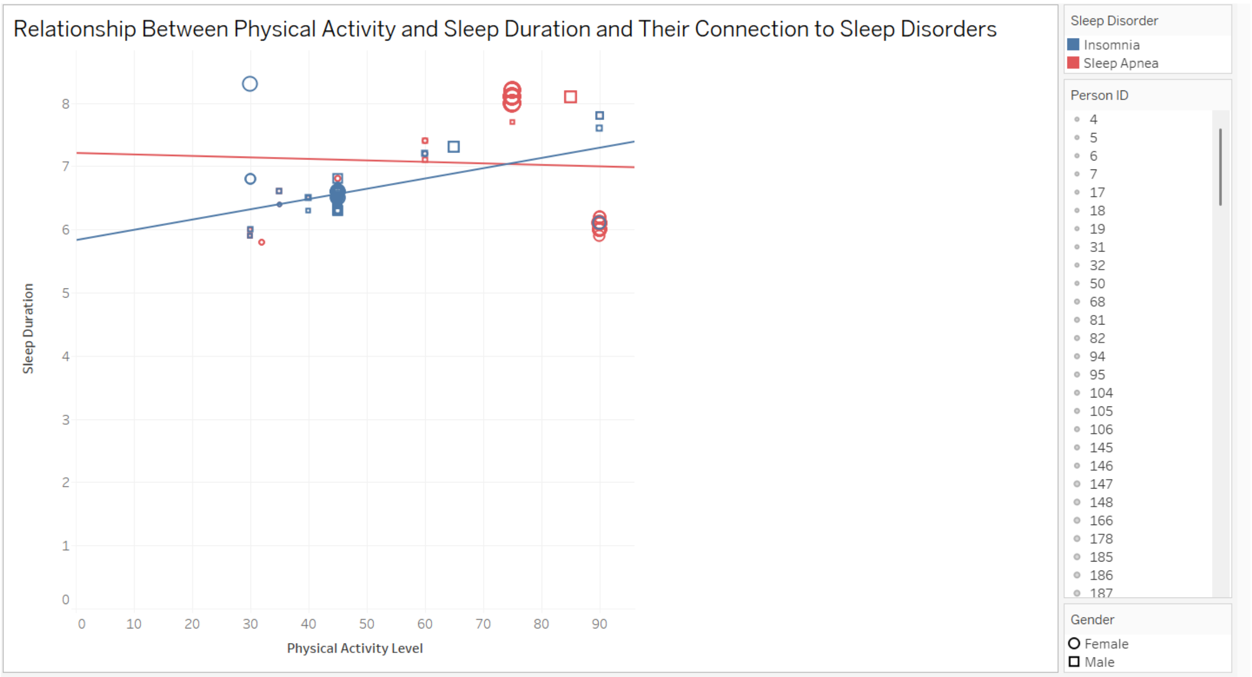


The stress analysis based on BMI categories and gender is presented in the scatter figure. The percentage of BMI categories—underweight, normal weight, overweight, and obese is expressed on the X-axis, and the stress levels amongst the members of the category are expressed on the Y-axis. Based on the plot, stress for both sexes escalated with BMI, but for each BMI category, female subjects recorded higher stress than male subjects in the outset BMI points. This means that, concerning stress, the results could be different and the differences could be observed when comparing women and men in terms of BMI. For the causative factors, further research is required for the data point with the highest stress level, which is an outlier. This figure shows the distribution of stress levels by BMI and raises attention to the possible gender differences.

Systolic pressure by the BMI categories is presented in scatters and subdivided by gender. The X-axis refers to the number of BMI categories while the total systolic pressure recorded for each of the categories' members is in the Y-axis. Thus, each data point has a gender-related coloring: orange is used for men, and blue, for women. In the plot, analyzing the difference in systolic pressure by gender as a function of BMI, we can observe that there are major differences in systolic pressure between boys and females in every explored group. They also report that the first BMI group's systolic pressure is higher among females than males, which means that there are some gender differences in the association of BMI with systolic pressure. As one can discover, this is the highest value of systolic pressure among all the data; therefore, additional scrutiny is needed to define the causes of such a rather striking outcome.

Same as in the case of males, the third scatter plot also shows the correlation between the systolic pressure and stress of the female participants in terms of trend lines showing a greater trend. Nevertheless, the interception and gradient points of the trendlines for performance differ based on gender, signifying the strengths. People with values of systolic pressure and perceived stress levels which are quite different from the other participants could be represented as cases, which are value deviations from the trend lines. This does, though, assert the relationship between stress levels, systolic pressure, and gender but the authors suggest requiring further research to find out the factors that cause these linkages because of the potentiality of high fluctuation and the existence of outliers.

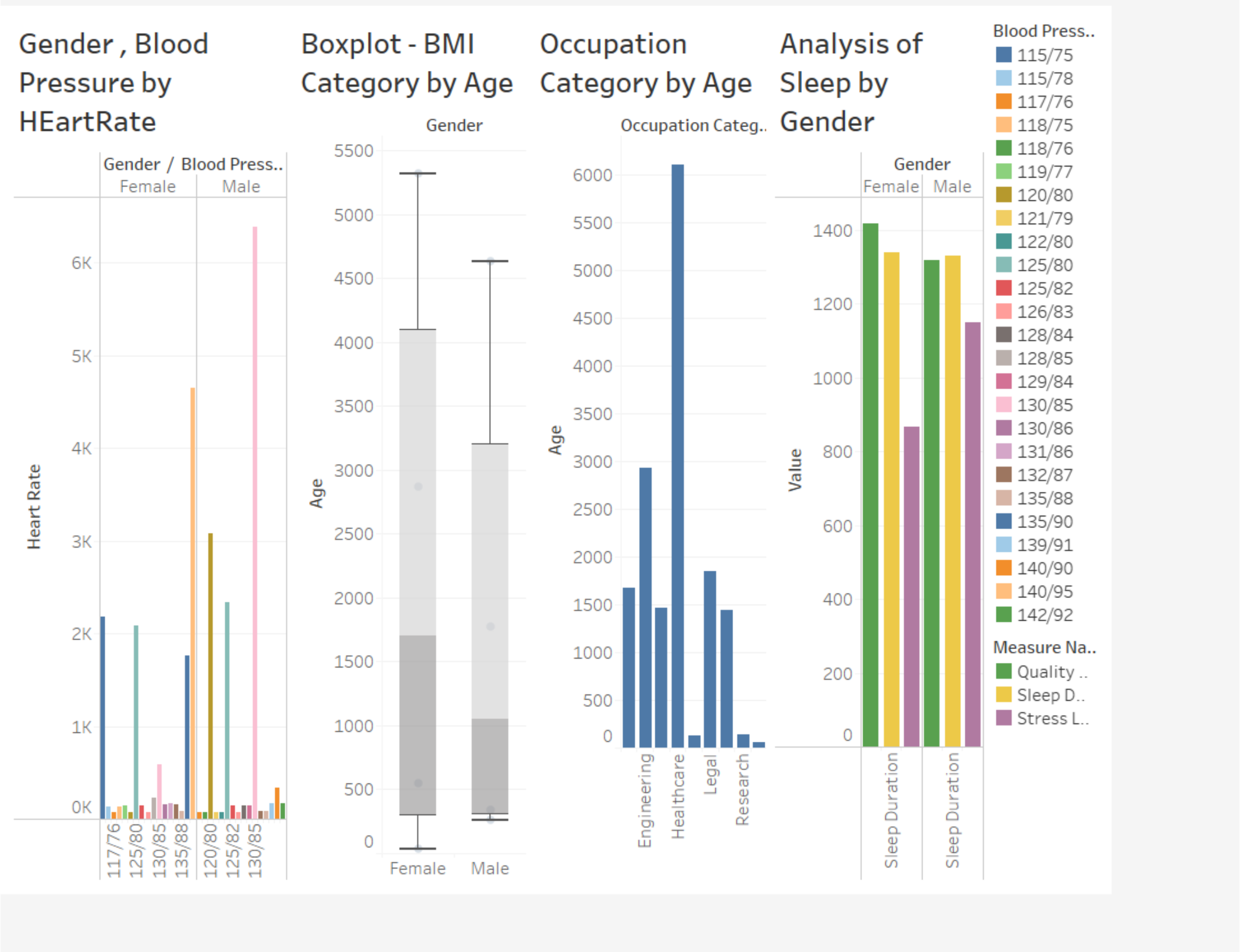
Task 4



The obtained data with the constructs on gender and sleep problem aspects are depicted in a scatter plot as the relationship between physical activity level and sleep length with trend lines. It is important to note that while labeling this graph, the primary X-axis represents the degree of physical activity performed while the Y-axis indicates the amount of sleep the said person gets. All the data points are gendered; the male is identified with a square while the female is represented with a circle and differentiated simultaneously by the type of sleep condition: sleep apnea (orange), and insomnia (blue). These trends revealed by the trend lines, depict the general trend of the data points with each of the categories of sleep disorders.

This graph shows possible interactions between gender and sleep issues affecting the relationship between physical activity and sleeping hours. Separate trend lines show that slope and intercept for each gender and sleep problem group are significantly different, and therefore the general tendency of sleep duration to increase with the growing physical activity level is observed. However, sex, and the type of sleep disturbances, make the level of their association vary. Notably, some of the observations fall off the lines indicating the possible presence of individuals with very specific sleeping or exercising habits. The scattering of these two data labels along these line inclinations offers a strategic analysis of the variability of the connection; in addition, visual learning is effective in highlighting patterns and peculiarities inherent in the data.

Task 5



The objective of Task 5 was to build four interactive visualizations that would help to perform the data investigation in different directions. By presenting related data on various pages and featuring dynamic data interaction, the dashboard takes a step further towards allowing even better navigation and data control to be performed by the users. The diagram represents age as a basic flow, the pressure of blood rate by heart rate, and the sleep of women and men, as well as also plots the card of BMI by age.

All the graphs show the parameter differences in sleep between genders such as sleep time, sleep efficiency, total movement, and wake after sleep. While clicking on any chosen data points or categories enclosed within a single visualization, users could additionally interact with the dashboard components, evaluating covariations between variables and looking for connections, patterns, or tendencies in the process.

Organized data-information exchange and decision-making processes become possible with the responsible approach to contemporary TV viewing. First of all, after processing a dataset the dashboard presents detailed information and lets users directly drill into the similarities, trends, and distribution within the data.

The figure is called “Labor Category by Age” and it looks like a bar chart. On the chart, the horizontal axis shows different age groups while the vertical axis displays the numbers of people in every labour category.

The age distribution charts by occupational category are good resources that one can use to know the makeup of the workforce. They can show things like:

Should we target young age groups or a general population group for specific occupations?

If it applies to a group of older people or younger people, this will be the type of employment or job. Whether this re-arrangement of workers among different fields will take place and how the future distribution varies from the current one.

The tabulated data in the picture present the blood pressure values for different genders expressed at different heart rates. On the negative side, excessive human measurements result in missing pulses, lack of blood pressure measurement (diastolic or systolic), and increased age which significantly influences atrial depolarization. Men substantially have a higher blood pressure than women, in particular aging men, but women are on are little more rapid heartbeat. The link between heart disease and risks like smoking, obesity, cardio inactivity, and family history has been established. Only when all relevant parameters are captured and investigated thoroughly can valid findings be reached going beyond the available evidence. Beyond this shortcoming, this picture leads to an understanding of the inequalities of both genders in the traits, specifically, heart rate, blood pressure, and pulse rate.

The "Analysis of Sleep by Gender" chart includes men's and women's average length, quality, and stress levels within sleep. Women's sleep is longer than that of men's, but women are more likely to have poor-quality sleep due to excessive stress. The table exhibits that the slumber routine amongst adolescents differs from either grownups or young children, with the latter ones going to bed later and waking up later than their age-mates.

Conclusion:

A detailed understanding of airline customers’ satisfaction and healthcare aspects could be derived by applying both Tableau data visualization and Python statistics analysis. Airlines use Python to obtain such data as guest evaluations to come up with a new path or strategy this is according to the passenger's wants and behaviors. Airlines can create a positive experience among customers and increase their loyalty by identifying feedback patterns obtained from users and through big data analytics such as sentiment analysis and classification algorithms.

On the other hand, Tableau visualizations permit to study of the relations between fitness indicators like arterial pressure, body mass, and stress level, as well as physical activity, sleep time, and sleeping disorders. Through the implementation of this information, the chance for airlines to make targeted measures on what will improve passengers' experience and overall well-being is increased. Health issues and customer satisfaction are gaining significance in the tough aviation sector; airlines can look into these concerns and can improve the feel of long-term loyalty and rewards.