**Solent University**

**in Partnership with QA (QAHE)**

**Coursework Assessment Brief – Apprentice**

**Assessment Details**

|  |  |
| --- | --- |
| Unit Title: | Analytics & Business Intelligence |
| Unit Code: | QHO539 |
| Unit Leader: | Dr Dhouha Kbaier |
| Level: | 5 |
| Assessment Title: | Personal Learning Record |
| Assessment Number: | AE1 |
| Assessment Type: | Personal Learning Record - PLR |
| Restrictions on Time/Word Count: | 2500 words |
| Consequence of not meeting time/word count limit: | There is no penalty for submitting below the word/count limit, but students should be aware that there is a risk they may not maximise their potential mark. |
| Individual/Group: | Individual |
| Assessment Weighting: | 100% |
| Issue Date: | September 2023 |
| Hand in Date: | **2nd February 2024 before 4 PM (local UK time)** |
| Planned Feedback Date: | 4 weeks after late submission deadline |
| Mode of Submission: | Online |
| Number of copies to be submitted: | One copy |
| Anonymous Marking | This assessment: is exempt from anonymous marking. |

Table Contents

[1. INTRODUCTION 3](#_Toc157365406)

[1.1 OVERVIEW 3](#_Toc157365407)

[WEEK 01 4](#_Toc157365408)

[WEEK 02 8](#_Toc157365409)

[WEEK 03 9](#_Toc157365410)

[3.1 DESCRIPTIVE STATISTICS 9](#_Toc157365411)

[WEEK 04 PANDAS- MATPLOTLIB - SEABORN 17](#_Toc157365412)

[4.1 VISUALIZATION. 17](#_Toc157365413)

[4.2 DATA PREPARATION 17](#_Toc157365414)

[WEEK 05 - WEEK 06 27](#_Toc157365415)

[TABLEAU VISUALIZATION 40](#_Toc157365416)

[REFERENCE 53](#_Toc157365417)

[APPENDICES 55](#_Toc157365418)

Alessandra Salanitri Francesca

ID: 10266448

Analytics and Business Intelligence  
QHO539

# INTRODUCTION

This report encompasses a diverse range of concepts, methodologies, and tools including weekly logs of taught concepts during the semester and, the Analysis of the Titanic dataset using Jupiter Notebook in Colab and, Tableau.

## OVERVIEW

For solid foundation, this report begins exploring different topics such as job position in Data Analysis, skills required and Data in Organizations.

It analyses then the Titanic dataset using Google Collaboratory. The analysis is divided into three files to simplify the readability. The three files are attached as objects in the appendices.

The first file, named: “First” demonstrates the ability to retrieve specific information from the dataset using Pandas library, and descriptive statistics by filtering the data based on column names and data types.

The second file, named: “Second” illustrates the process of pre-processing data for visualization using Pandas, Matplotlib, and Seaborn libraries.

The third file, named: “Third” uses correlation, linear regression, train, and test split tools to extrapolate further insight from the data.

Finally, Tableau is used to present visualizations that effectively communicate key insights derived from the Titanic dataset.

# WEEK 01

The aim of the first weekly log in this module is to explain the different roles in the dynamic landscape of the data-driven decision-making domain.

**Statisticians**, **Data Analysts**, and **Business Analyst** are professionals who work with data that is then utilized for insights and decision-making with different focuses and responsibilities.

The differences between these roles, responsibilities, and skills required for these professions are illustrated in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Role** | **Responsibilities** | **Skills** |
| **Statistician** | Expert in collecting, analyzing, interpreting, presenting, and organizing data | Design surveys and experiments, collect and process data. Data-driven decisions are crucial. | Mathematical and statistical skills, knowledge of statistical software, draw meaningful conclusions from data. |
| **Data Analyst** | Expert in interpreting and analyzing complex data sets | Clean and process data, perform data analysis and create visualizations to help organizations make informed decisions. Identify trends, patterns, and insights from data. | Knowledge of programming languages, data visualization tools, and databases. Analytical thinking. |
| **Business Analyst** | A bridge between business needs and solutions. Use data to improve efficiency, identify opportunities, and solve business problems. | Work along with stakeholders. Focus on statistical analysis and also use data to inform strategic and operational decisions. | Analytical skills, business acumen, and communication skills are essential as needed to translate technical information into business terms. |

In summary, statisticians specialize in statistical methodologies, data analysts interpret and analyse data, and business analysts bridge business needs with data-driven insights.

The data scientist offers Domain Knowledge, crucial to interpreting results in a meaningful context or selecting relevant features. The skills required for becoming a data scientist include a combination of technical, analytical, and domain-specific skills as:

Proficiency in programming languages such as Python, R, and SQL is often essential to sorting, analysing, and managing big datasets.

Statistics and probability, which enable individuals to write high-quality machine learning models and algorithms.

Data wrangling and database management: where the data is processed, cleaned, and organized.

Machine learning and deep learning help to gather and synthesize data efficiently, predicting outcomes of future data sets, using linear regression, logistic regression.

Data Visualization: technologies such as Tableau, Microsoft Excel, and Power BI for graphs and charts creations, can help to present a clear visualization of the business insights in a friendly manner.

Interpersonal skills such as Active listening, Effective communication skills, sharing feedback, Attention to detail, Leadership, Empathy, and Public speaking help to form strong working relationships with team members and present findings to stakeholders.

To ensure a successful integration of a data scientist into its operation, an organization is required to meet essential requisites such as

give access to a substantial amount of quality data to allow for robust and accurate model training,

offering adequate data storage solutions necessary to store and manage large datasets and, appropriate resources for processing and analyzing data efficiently,

clear Problem Definition which helps data scientists choose appropriate algorithms, metrics, and methodologies for solving specific problems,

Data Governance and Compliance to ensure data scientists adhere to privacy regulations and data security measures are in place to protect sensitive information.

Business Intelligence (BI) is a broad term that encompasses data mining, process analysis, performance benchmarking, and descriptive analytics. Business intelligence plays a pivotal role in helping organizations to gain insights through historical, current, and predictive views of business operations, contributing to their success and competitiveness.

By analysing customer's behaviours, businesses can customize products and services to increase loyalty and satisfaction.

By leveraging BI effectively, organizations can respond to market changes faster and identify opportunities quickly. A BI system allows real-time monitoring of KPIs so that issues can be addressed promptly as they arise. BI contributes to cost reduction by identifying inefficiencies and optimizing processes.

It can also contribute to revenue growth by identifying new market opportunities, improving sales and marketing strategies, and enhancing customer relationships while supporting risk management by identifying potential risks.

Below are three job roles available online in BI indicating the required skills, salary, and job benefits. Links are attached to the reference. (PAGE 53)

|  |  |  |
| --- | --- | --- |
| **Job Title** | **Job description** | **Salary** |
| *Business Intelligence Analyst* | In this role, you will analyze and produce reports on risk and performance across different complex datasets. You will regularly present and communicate business performance data to help the organization understand its performance, celebrate success, and address any issues. In addition, you will look to actively improve how we use systems and dashboards, working with internal stakeholders to improve reporting and data accuracy e.g. through Microsoft Power BI and other visual presentation tools. | 30.350£ |
| *Business  Intelligence  Team Lead* | Provide knowledge and experience to the company to best solve problems and action upon opportunities in an insurance environment. To provide detailed data-based analysis of Business processes and scenarios. To manage and develop the team of technical analysts, ensure informed decisions are made and new methodologies are implemented to help achieve the Group's strategic objectives. Help shape and determine a reporting strategy, including an understanding of available technologies and software to assist delivery. Ability to make considered decisions in a pressured environment, ensuring understanding and impact of consequences is understood. | 60.000£ |
| *Business Intelligence  Manager* | Managing a team of 3 from junior to senior level including Analysts and Engineers so previous experience is desirable. Rolling out self-serve dashboards to the wider business. Using SQL and data visualization tools to create reports and dashboards for the business. Communicating and gathering requirements from stakeholders at all levels. Developing and executing the BI strategy aligned with business objectives | 75.000£ |

In conclusion, each role demands specific skills, with corresponding salary increases reflecting the increased responsibilities. For further details, appendix 01 displays Business Analyst Salaries based on Experience.

# WEEK 02

Using the web and online resources, the second week aims to analyze the different organizations based on the industry, services offered and what type of data can be retrieved from those organizations. Below, three organizations are explored.

Week02\_PLR file is embedded to display further information on other organizations.

A table of information

Description automatically generated

# WEEK 03

## DESCRIPTIVE STATISTICS - PANDAS

This paragraph begins with analysing the Titanic dataset in Jupiter Notebook.

In “First” file, descriptive statistics are used for acknowledging some general and some detailed information about the Titanic dataset used from Seaborn library.

A screenshot of a computer program

Description automatically generated

“Pandas was imported as pd”, where pd is the short name indicating that the Pandas library is being called.

Pandas stand for “Python Data Analysis Library”, it is an open-source Python library used for analysing, cleaning, exploring, and manipulating data.

“math” is a basic math python library.

“Numpy was imported as np”, where “np” is the short name indicating that the numpy library is being called. Numpy is an open-source Python library, used for numerical data in Python.

“Matplotlib.pyplot” was imported as “plt”, where “plt” is the short name indicating that the matplotlib.pyplot library is being called.

Matplotlib is a Python library for create 2D graphs and plots by using Python syntax.

The “pyplot” module provides features to control graph proprieties as title, axes, etc.

“Seaborn” library as “sns”, where “sns” stands for Seaborn. Seaborn is a Python library used for high-level statistical graphs.

Additionally, %matplotlib allows your graphs to be included in the notebook file, next to the code.

As an initial prompt, the info() method is used to provide a comprehensive overview of the dataset, including the total number of entries, data column names, the total number of columns, data types - indicating whether the data is categorical or numerical-, and the presence of null values. Null values represent missing data points and can be the cause of challenges in data analysis as it can change the results of the analysis based on unreliable data.

A screenshot of a computer

Description automatically generated

The head() method displays by default, the first 5 rows of the dataset.  
The query in the code below requests specifically the first 10 rows of the dataset, this is accomplished by inserting the desired value inside the parenthesis.

This method offers a first basic insight into the data structure and initial observations.

A screenshot of a computer

Description automatically generated

To displays one, two, or more specific columns allowing further exploration of specific data points, the code below is used.

In this case, because it is not specified how many rows are being required, by default, the first 5 rows will be displayed.

A screenshot of a computer

Description automatically generated

The tail() method, displays by default the last 5 elements of the dataset.

In the code below, however, the query is specifically requiring the last 10 rows.

A screenshot of a computer

Description automatically generated

The loc[ ] method retrieves specific rows based on the index value provided inside the brackets. In this case, the query extracts data from rows 10 to 20, displaying a targeted view of the dataset.

A screenshot of a computer

Description automatically generated

The shape() method provides a concise representation of the dataset's dimensions, returning a tuple containing the number of rows and columns.

A close-up of a number

Description automatically generated

The .dtypes attribute explicitly specifies the data types of each column, enabling a clear understanding of the data's structure.

A screenshot of a computer

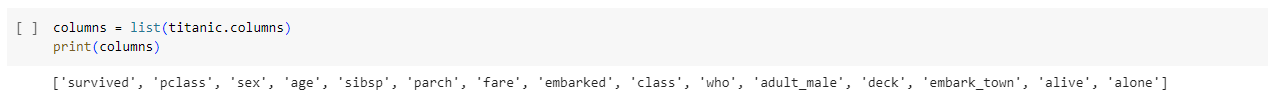
Description automatically generated

The .columns attribute retrieves the dataset's header information in the form of a Pandas Series.

A white background with black text

Description automatically generated

The list() function converts this series into a standard Python list for further manipulation.



The .describe() method generates comprehensive statistical summaries of the dataset, including:

A count of null values for each column,

Mean, standard deviation (std), minimum, and maximum values for numerical columns.

A screenshot of a computer screen

Description automatically generated

As shown in the code below, by specifying a column title within the brackets, the mean for a particular column can be isolated and displayed.

A screenshot of a computer

Description automatically generated

Additionally, formatted strings can be employed to limit the decimal places of float numbers using the xf format, where x represents the desired decimal precision as shown in below code lines. This ensures a better readability for stakeholders and decision maker.

A screenshot of a computer code

Description automatically generated

The. select\_dtypes() method enables the extraction of columns based on their data types. The exclude and include parameters allow for selective filtering, making it easier to isolate specific data subsets.

The line of code displayed below is querying the categorical type of data.

The result illustrates the top categorical value used.

A screenshot of a computer code

Description automatically generated

The line of code below is querying the numerical type of data. Allowing for further analysation options throughout statistics.

A screenshot of a computer

Description automatically generated

The code below is used to find the mean() , also known as average.  
The mean value is generally used for statistics and data analysis. It summarizes the entire dataset with a number that represents the data’s centre point.

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

The mode is the most frequent value stored within a column.

It is the only average value that can have no value, one value or more than one value.

One value because if none of the number are repeated, that column does not have a mode. More than one value because if there are two values in the same column to have the same frequency, both values are the mode.

Mode is often used in sample. Sample is a dataset collected from a targeted population.

A screenshot of a computer

Description automatically generated

Median is the value in the exact middle of the data.

To calculate the median, the value in the column needs to be sorted from smallest to largest.

A computer screen shot of a computer

Description automatically generated

The standard deviation or std measures the variation or dispersion in a set of values in a single number. It helps to identify trends and assess data reliability.

A black text on a white background

Description automatically generated

Variance measure how far the outliers are spread far from the mean.

A close-up of a number

Description automatically generated

-“Skewness is a measurement of the distortion of symmetrical [distribution](https://www.investopedia.com/terms/d/distribution.asp) or asymmetry in a data set.”- Ref. 4th.

The distribution of the data is illustrated as a curve. When the data do not have a normal distribution and, the curve is skewed to the left side, this is called a negative skewness. When the curve is skewed to the right side it is called the positive skewness.

A screenshot of a computer

Description automatically generated

A close-up of a code

Description automatically generated

Overall, this first file demonstrates the versatility of Pandas for data manipulation and analysis. The ability to extract, filter, and summarize data effectively facilitates deeper insights into the Titanic dataset.

# 04 PANDAS- MATPLOTLIB - SEABORN

## VISUALIZATION.

The initial step of “Second” file is importing all the libraries necessary to extrapolate the information required to create visualizations.

A screenshot of a computer code

Description automatically generated

## PRE-PROCESSING

For plotting and visualization, it is required that the dataset has only reliable and usable data, hence, all duplicates’ values, null values needs to be dropped.

The code below queries and display all the null values and the duplicates in the Titanic dataset.

A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

The code below queries to display a raw dataset, then queries to drop duplicate and null values and to display the new information with the clean dataset. The inplace = True, enable the data to be renamed in place.

A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

To present an understandable report to Stakeholders and decision maker, our data needs to be as clear as possible. In this dataset, for example, some titles are not clear or descriptive such as sibsp, parch, etc.

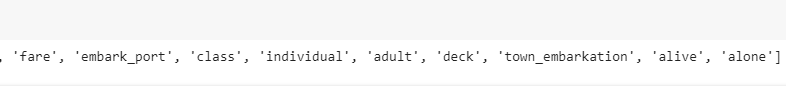
Below code display columns titles being renamed with descriptive title replacing for example “sibs” title with “siblings or spouse” title.

A screen shot of a computer code

Description automatically generated

A close-up of a white background

Description automatically generated



Now, the dataset has only reliable, understandable and usable data and it is ready to be analysed through graphs.

Exploring the Titanic dataset reveals many valuable information about the tragic disaster happened.

The insights revealed in dept from titanic dataset are displayed in an understandable manner thanks to different graphs following.

Count plot counts the number of observations in each category and displays it as a bar chart.

Illustrating below the number of individuals who managed to survive.

A screenshot of a graph

Description automatically generated

This graph displays a title:” Survived and Not survived”, a label indicating the survived (x axis), while the other axis represents the count of individuals.

The legend is displayed thanks to the voice “hue” which act to separate the visualization of not survivors in blue colour, and the survivors in orange colour.

The graph suggests that around 122 people survived, while around 59 people died in the ocean.

The next graph analyses the rate of survival between female and male individuals.

It suggests that the proportion of

female survived (82) is higher than

males’ survival rate (41).

The mortality rate for female is lower (7) compared to males (54). This could be related to “women and children first” policy during emergencies. Further information could be found analysing this variable in dept.

A graph of a person

Description automatically generated

The next graph analyses the passenger class and the survival rate.

The graph shows an exponential difference of survival rate based on the class each passenger was traveling in. The first class displays a higher rate of survival, while the second- and third-class passengers had low to none rate of survival.

The commented line #plt.ylabel (‘count’), assign the title “count” to “y axis”, however, because the action of the command is to count, the world “count” is displayed by default.

A graph of a number of people

Description automatically generated with medium confidence

The next graph analyses the “Travel Status of Passengers” considering either passengers were travelling alone or with families’ members displayed in a vertical bar chart.

A screenshot of a computer screen

Description automatically generated

In the next graph, the Age Distribution by Passenger Class is displayed using box plot.

Box plot is used to display the minimum, the maximum, the median and outliers of the age variable among each passenger class.

The graph helps to easy compare the tendency and potential outliers within each class and could help identify patterns or differences in age distribution.

Below graph shows the passengers travelling in first class had an average of 38 years old, with the minimum number being 1 year old and the maximum being 80 years old. The second-class passengers had an average of 30 years old, with the minimum age being 2 years old and the maximum being 58 years old. The third class passengers had an average of 25 years old, with the minimum age being 3 years old and the maximum of 42 years old.

A screen shot of a graph

Description automatically generated

The graph below shows the relationship between variable “age” and “variable” fare.

It is suggested a strong correlation between the two data points.

A screen shot of a graph

Description automatically generated

The next graph displays the age distribution within classes using a line plot.

Line plot graph uses data points displaying a clear representation of trends and patterns.

A screen shot of a computer

Description automatically generated

# SKLEARN

## TRAIN/TEST SPLIT

In “Third” file, sklearn library is imported along Panda, Seaborn, Matplotlib and numpy.

Sklearn is a Python library widely used to implements pre-processing, cross-validation, regression, visualization, and algorithms.

From “sklearn.model\_selection” module , the function “train\_test\_split”, splits arrays into random train and test subsets to wrap input validation.

From “sklearn.linear\_model” module, the class “LinearRegression”. This class is used to perform linear regression, a basic predictive analytics technique. This technique is illustrated within this report (PAGE 34).

From “sklearn.linear\_metrics” module, the function “mean\_squared\_error”.

Mean square error (MSE) is a measure of how close a predicted value is to the actual value.

The imported module “warnings” is part of the Python standard library. The “warnings.filterwarnings() line, sets up filters for warning messages. In this file, it sets the ‘ignore’ action, hence any warning messages occurring during the program’s execution will not be printed to the console. In real life, warning messages are helpful because alert the developer regarding issues that may affect the program.

A screenshot of a computer code

Description automatically generated

The classes imported from scikit-learn, are used to preprocess: “adult, alone, gender, alive and town embarkation” columns which are originally categorical data.

Sklearn preprocessing module “Label Encoder” class, convert categorical data into numerical data. “OneHotEncoder” class convert each categorical integer into one-hot encoded new binary columns. Sklearn compose module “ColumnTransformer” class, used to preprocess datasets with a mix of categorical and numerical values. It helps by applying different transformations to different subset of column types.

A close-up of a white background

Description automatically generated

“Le”, stands for label encoder.

A screenshot of a computer code

Description automatically generated

“Drop duplicates()” translate for each column, the categorical data into unique numerical values as 0 or 1.

The “fit” method is used to fit each label encoder on the unique values of the corresponding categorical column.

A screen shot of a computer code

Description automatically generated

“Transform” convert the actual values in the categorical columns to their corresponding numerical labels and set the new columns with the current titles plus “enc”.

A screenshot of a computer code

Description automatically generated

Below illustrated the full code and result

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In below code, the remainder is initialized with the “ohe” alias which stand for OneHotEncoder, the transformer will be applied in the “town\_embarkation” column. The “passthrough” parameter indicates the column which have not been called for any transformation process should be kept as it is.

A screenshot of a computer program

Description automatically generated

In the code below, a new Data Frame named “ins\_titanic” is created using the transformed data. The “ct.get\_feature\_names\_out” method, return the name of the transformed data. The “list(ins\_titanic.columns)” convert the column names of “ins\_titanic” to a list and prints the results of all the transformed data and the ones that was kept without transformation.

A screenshot of a computer program

Description automatically generated

As displayed below now each value of “town\_embarkation” now has its unique column and has a value of 1 for existing individual embarked from that specific town and, a value of 0 for a non-existing individual embarked from that specific town.

A screenshot of a computer

Description automatically generated

A white background with many lines

Description automatically generated with medium confidence

A white background with many small colored lines

Description automatically generated with medium confidence

In the code below the column titles in the “ins\_titanic” dataset are being renamed as specified in the list.

A screenshot of a computer code

Description automatically generated

In the code below, the columns are reordered.

A screenshot of a computer code

Description automatically generated

Finally, a new data frame “ins\_titanic\_2” is created with selected columns from the existing data frame “ins\_titanic”. Object columns are being removed.

The “pd.to\_numeric” convert the column passed by the list, to numeric data types.

A screenshot of a computer code

Description automatically generated

The info() and head() method displayed all columns types are either float or integer hence, numerical value.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The code below request, for the item specified in the list, the correlation calculation.

The method “.corr()” calculate the correlation coefficients between each selected columns and store them in a new data frame named “df\_corr”.

If the correlation coefficient is 1 or close to 1, indicate a good or strong correlation.

If the correlation coefficient is -1 or close to -1, indicate a negative or no correlation.

A screenshot of a computer

Description automatically generated

Below displayed a heatmap, which provides a visual representation of the correlation matrix previously analysed. On the right-hand side of the map, the legend illustrates how the colour intensity indicates the strength of the correlation. This helps quickly identify relationship between the data.

A screenshot of a computer

Description automatically generated

The code below uses a subset of values from the dataset for training and testing purposes.

A screenshot of a computer program

Description automatically generated

This line of code : “” select all rows and columns except the last one as X.

This line of code : “” select all rows and the last column as Y.

This line of code : “ ” splits the dataset into training and testing. The result is stored into the variables : x\_train, x\_test, y\_train, y\_test. The random\_state=54 sets a random seed for reproducibility and, test\_size specifies how much of the data will be used for testing and how much for training. In this case, 30% (0.3) of the data will be used for testing and, 70% for training.

Thanks to this code below, the “model” has learned the relationship between the feature x\_train and the variable y\_train based on the linear regression algorithm. This model can now be used to make predictions or evaluation on its performance on a testing set.

A screenshot of a computer program

Description automatically generated

Below illustration on how to extract the intercept “a” and coefficient “b” from the LinearRegression model previously trained.

The “model.intercept” contain the intercept “a” in the LinearRegression model. It represents the predicted value.

a = model.intercept\_

The coefficient “b” represents the change in the predicted value.

B = model.coef\_

A screenshot of a computer program

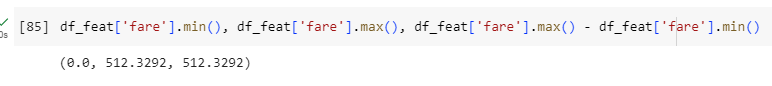
Description automatically generated

A screenshot of a computer code

Description automatically generated

These metrics provide MSE (mean squared error), RMSE (square root of MSE), MAE (mean absolute error) and R-square, which measure the proportion of the variance in the predictable variable from the independent variable.

Below code calculate minimum, maximum and the difference between, of the “fare” column. This process can provide insight on the distribution of the “fare” within the dataset



Finally displayed below is the model prediction.

The model takes as first argument the passenger class which can either be 1,2,3, the gender (0 for woman, 1 for man), the age, either the passenger will be traveling with siblings or spouse (insert number), either the passenger will travel with parent or children (insert numb), either the passenger will travel alone (0 for no, 1 for yes).

Once all the arguments have been inserting on the model, it will predict the price of the ticket for this passenger.

This example takes a passenger requiring 1st class, woman, 20 years old, traveling with 3 passengers, 0 parents or children, not alone. The predicted fare for this passenger is: 122.73£

A computer code with numbers and symbols

Description automatically generated with medium confidence

This example takes the same details as the previous example, changing only the gender. The predicted fare for this passenger is: 107.20£

A screen shot of a computer code

Description automatically generated

This example takes the details of a 40 years old, woman, in 1st class travelling alone. The predicted fare is: 66.21£

A computer code with text

Description automatically generated with medium confidence

This example takes the details of a 1st class passenger, male, 40 years old, traveling with a spouse or siblings, three children or parents, not alone. The predicted fare is: 166.26£

A computer code with numbers and symbols

Description automatically generated with medium confidence

More example are displayed below. Now, for every new passenger request, it is possible to use the model to generate a prediction for their ticket fare.

A screenshot of a computer program

Description automatically generated

# TABLEAU VISUALIZATION

Tableau is a BI tool is used to create sophisticated and easily understandable data visualization. Thanks to this tool we can display self-explanatory visualizations about the incident occurred in April 1912 in Titanic.

Three different datasets are used for a better understanding.

For the first worksheet, the Details csv file dataset is used to display the incident Route map which suggests where the Titanic was Stationed and departed, where it sank, and where its destination was.

A screenshot of a computer

Description automatically generated

When the dataset was imported, the latitude column was not recognised from Tableau as a geographical role. Hence, it was changed manually as displayed below.

A screenshot of a computer

Description automatically generated

The second worksheet is the Passenger Provenience, using the csv dataset Titanic3.

A screenshot of a computer

Description automatically generated

Below image, displays a Dashboard with four different graphs analysing some insight from the Titanic Data.csv which displays important details regarding passenger distribution by class, age by class, survival by class and survival by gender. Tableau file is attached as object in the reference, and it stores individual graphs used for this Dashboard.

A screenshot of a computer screen

Description automatically generated

Below is a dashboard containing some images and pie charts, which analyse the status of survival for child, female and male. Individual analysis are stored in Tableau file attached in the references.

A screenshot of a computer

Description automatically generated

Some step demonstrating the execution are displayed below:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The percentage is displayed thanks to the table calculation created on the CNT titanic dataset – which is an autogenerated value from Tableau which contains the count of the dataset.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

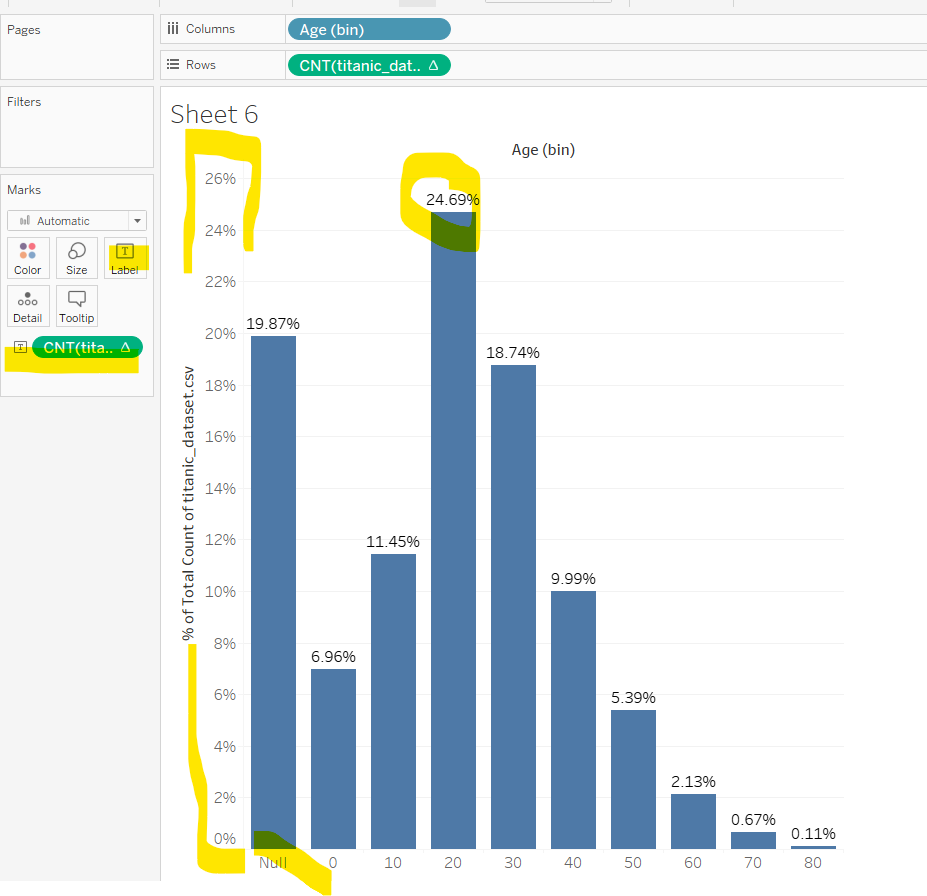
A screenshot of a computer

Description automatically generated

Creating a calculation field to recognise child and adult in the age calculation value.A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated



Tried to insert some linear regression but dataset used was not good for linear representation. However, the step to use:

A yellow pencil on a white background

Description automatically generated

# REFERENCE

27/10/2023

Reference- This ref may be changed or updated at any time.

|  |  |
| --- | --- |
| **Title** | **Link** |
| **What do statisticians do? Roles, Responsibilities and Career Paths:** | <https://graduate.northeastern.edu/resources/what-do-statisticians-do/> |
| **Title: The role of a business Analyst – Key responsibilities and skills** | <https://lset.uk/blog/the-role-of-a-business-analyst-key-responsibilities-and-skills/> |
| ***What skills does a Data Analyst need to get hired?*** | <https://www.dataquest.io/blog/data-analyst-skills/> |
| **7 Skills Every Data Scientist Should Have** | <https://www.coursera.org/articles/data-scientist-skills> |
| **How to Hire a Data Scientist in 2023: Sourcing + Interview Advice** | <https://www.celential.ai/blog/how-to-hire-a-data-scientist/> |
| **What Is Business Intelligence (BI)? Types, Benefits, and Examples** | <https://www.investopedia.com/terms/b/business-intelligence-bi.asp> |
| **Job Search: Business Intelligence Analyst; Business Intelligence Team Lead** | <https://uk.indeed.com/q-business-intelligence-jobs.html?vjk=5ca7b37b4d4a93d3>  <https://www.bing.com/jobs?q=business+inteligence+jobs+uk&scp=0&jid=1725657017.Retro&rb=0&rc=20&L2=true&c=1&form=JOBL2S>  <https://www.bing.com/jobs?q=business+inteligence+jobs+uk&scp=0&jid=1725657017.Retro&rb=0&rc=20&L2=true&c=1&form=JOBL2S> |
| **Image appendix**  SKLEARN.MODEL\_SELECTION.TRAIN\_TEST\_SPLIT  MATPLOT LIBRARY  SEABORN LIBRARY  MEAN DEFINITION  STD DEFINITION  SKEWNESS | <https://www.bing.com/images/search?view=detailV2&ccid=XpJiMcdj&id=41CE248CE4E4BE3BB162D6E98BB4EEB28AF9FD04&thid=OIP.XpJiMcdjFIHh3rDfcwT53QHaDt&mediaurl=https%3a%2f%2fdpbnri2zg3lc2.cloudfront.net%2fen%2fwp-content%2fuploads%2f2021%2f07%2fBusiness-Analyst-Salary-2.jpg&cdnurl=https%3a%2f%2fth.bing.com%2fth%2fid%2fR.5e926231c7631481e1deb0df7304f9dd%3frik%3dBP35irLutIvp1g%26pid%3dImgRaw%26r%3d0&exph=597&expw=1190&q=ABI+salary+in+uk+from+starting+salary+to+experienced+graph&simid=608037193555803797&FORM=IRPRST&ck=BEE94A4E610E82FBDA2ADF76D0383828&selectedIndex=2&itb=0&ajaxhist=0&ajaxserp=0>  <https://scikit-learn.org/stable/getting_started.html>  <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>  <https://matplotlib.org/stable/api/pyplot_summary.html>  <https://seaborn.pydata.org/>  <https://statisticsbyjim.com/basics/mean_average/>  <https://statisticseasily.com/what-is-standard-deviation-statistics/>  <https://www.investopedia.com/terms/s/skewness.asp#:~:text=Skewness%20is%20demonstrated%20on%20a%20bell> |

# APPENDICES



