Coding Temple

Flex Data Analytics Program

R Project

Presentation

Spaceship Titanic Dataset (Kaggle)

Submitted by: Farrukh Sultan

Summary Statistics

Mean, Median, Min, Max, etc.) for numeric variables like Age and TotalBilling

- **For the Age variable:**

Minimum (Min.): The minimum age in the dataset is 0, which indicates the "Infants"

1st Quartile (1st Qu.): 25% of the passengers have an age of 20 or below.

Median: The median age is 27, which means that 50% of the passengers are 27 years old or younger.

Mean: The mean age is 28.77, which is slightly higher than the median, suggesting that the distribution might be slightly skewed to the right.

3rd Quartile (3rd Qu.): 75% of the passengers have an age of 37 or below.

Maximum (Max.): The maximum age in the dataset is 79, indicating the oldest passenger in the dataset.

- **For the TotalBilling variable:**

Minimum (Min.): The minimum total billing amount is 0, which might indicate that some passengers did not spend any money on amenities.

1st Quartile (1st Qu.): 25% of the passengers have a total billing amount of 0 or below.

Median: The median total billing amount is 716, indicating that 50% of the passengers spent 716 units or less on amenities.

Mean: The mean total billing amount is 1433, which is higher than the median, suggesting that the distribution might be right-skewed.

3rd Quartile (3rd Qu.): 75% of the passengers have a total billing amount of 1442 or below.

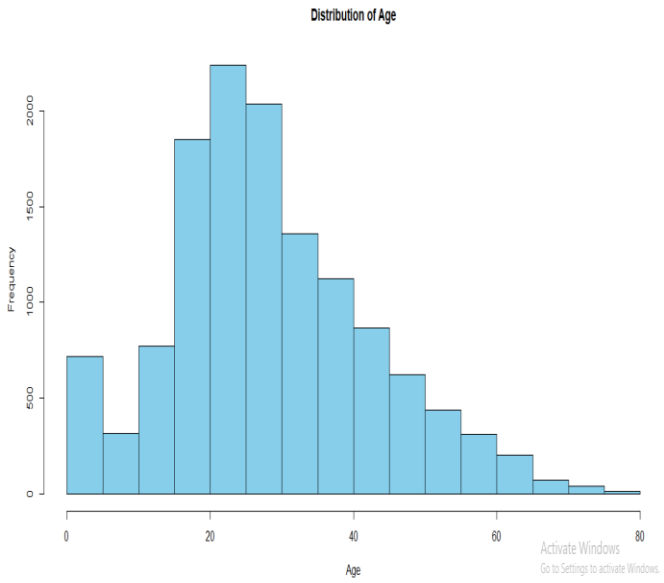
Maximum (Max.): The maximum total billing amount is 35987, indicating the highest amount spent by a passenger on amenities.

Frequency of VIP passengers

Majority of passengers (12,697) are not VIPs, while a smaller number (273) are VIPs.

Data Distribution

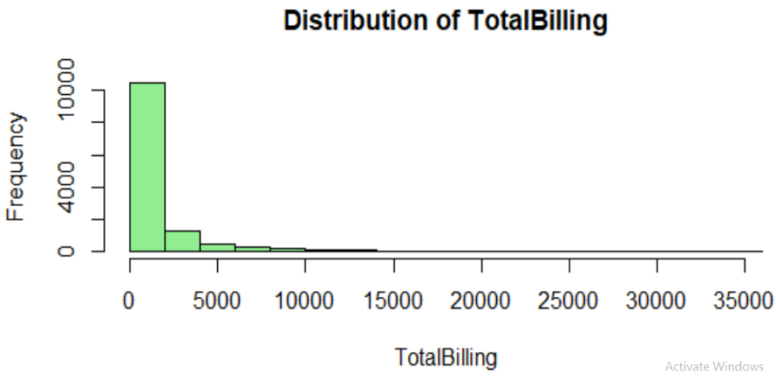
Distribution of Age using a histogram



Age distribution histogram shows a slowly sloping right side and an inconsistently sloping left side, it suggests an asymmetric distribution with potential outliers or unusual patterns at the lower end of the age range. The slowly sloping right side of the histogram indicates that there are fewer passengers at older ages compared to younger ages. This could be expected in many populations, as the number of individuals typically decreases with increasing age. The inconsistently sloping left side suggests that there may be irregularities or anomalies in the distribution of ages at the lower end. This could be due to factors such as assumption of infants, data entry errors, or outliers that skew the distribution.

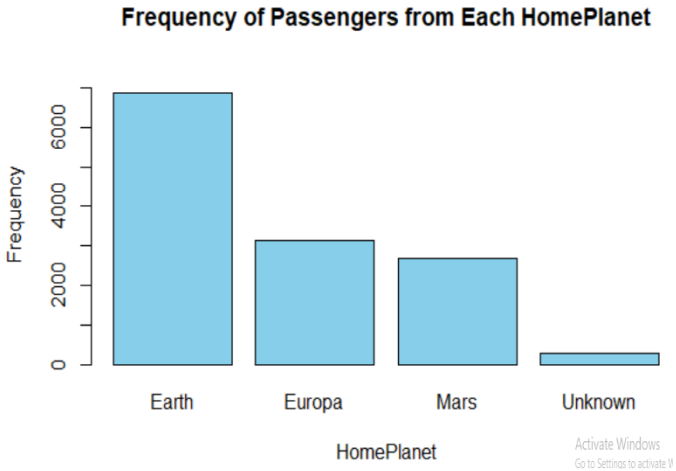
Distribution of TotalBilling to see the spending patterns of passengers

The distribution suggests that the majority of passengers have relatively low total billing amounts, indicating that they did not spend much on additional services or amenities during their voyage.

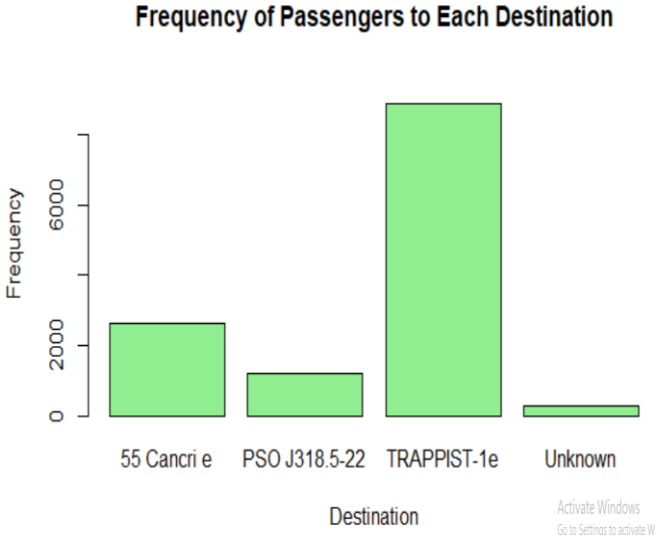


Categorical Analysis

Frequency of passengers from each HomePlanet and Destination using bar plots



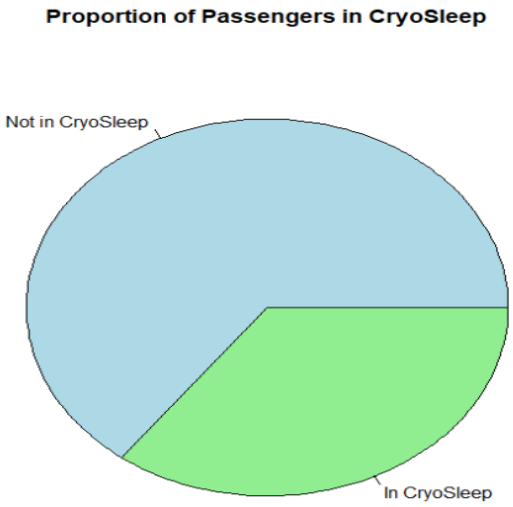
Earth has the highest frequency of passengers among all HomePlanets. This indicates that a significant portion of passengers in the dataset originate from Earth. It may imply that Earth is either the most populous HomePlanet or the most frequently traveled from by passengers in the dataset. Europa has the second-highest frequency of passengers after Earth. Mars has a lower frequency of passengers compared to Earth and Europa but still has a noticeable presence. The box for "Unknown" HomePlanet is the smallest, indicating the lowest frequency of passengers among the known HomePlanets. "Unknown" likely represents cases where the HomePlanet information was missing or unavailable in the dataset. While there are fewer passengers with "Unknown" HomePlanet, it's essential to note that this category may contain valuable information, such as passengers from newly discovered or unnamed planets or those with incomplete records.

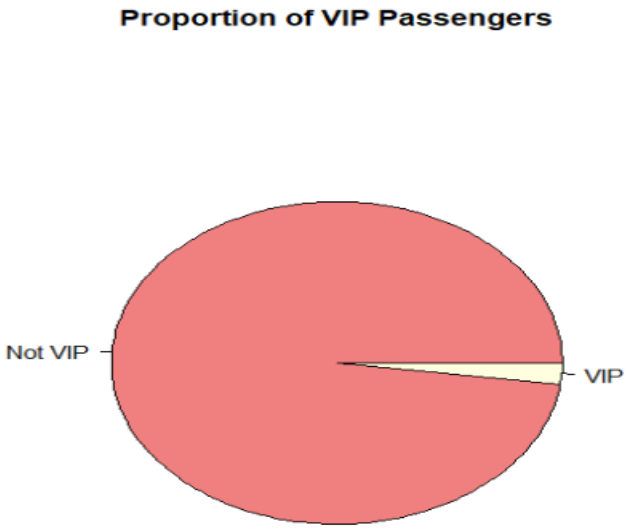


Trappist-1e has the highest frequency of passengers among all destinations. This indicates that a significant portion of passengers in the dataset are traveling to Trappist-1e. Trappist-1e may be a popular destination for interstellar travel in the dataset, possibly due to its habitable conditions or other factors. 55 Cancri e has a lower frequency of passengers compared to Trappist-1e but still has a notable presence. This suggests that 55 Cancri e is another destination of interest for passengers in the dataset, although less popular than Trappist-1e. PSO J318.5-22 has a lower frequency of passengers compared to both Trappist-1e and 55 Cancri e. While not as high as the previous destinations, the frequency of passengers to PSO J318.5-22 is still noteworthy. This indicates that PSO J318.5-22 is a destination that some passengers in the dataset are traveling to, although it may not be as popular as Trappist-1e or 55 Cancri e. The box for "Unknown" destination is the smallest, indicating the lowest frequency of passengers among the known destinations. "Unknown" likely represents cases where the destination information was missing or unavailable in the dataset. While there are fewer passengers with "Unknown" destination, it's essential to note that this category may contain valuable information, such as passengers traveling to newly discovered or unnamed destinations or those with incomplete records.

Proportion of passengers in CryoSleep and VIP status using pie charts

It indicates that a larger proportion of passengers have opted not to undergo CryoSleep during the journey. A higher proportion of passengers not in CryoSleep suggests diversity in preferences, circumstances, and choices among passengers aboard the spaceship. Understanding these variations can provide insights into the dynamics of interstellar travel and the factors influencing passengers' decisions during the journey.





A higher proportion of passengers opting not to obtain VIP service indicates that the majority of passengers are satisfied with standard service offerings and do not see the need for additional amenities or privileges. This could imply that the standard services provided onboard the spaceship are sufficient for meeting passengers' needs and preferences. The decision not to obtain VIP service may be influenced by economic factors, such as cost constraints or budget considerations. Passengers may prioritize cost-effectiveness and opt for standard service to minimize expenses during the journey.

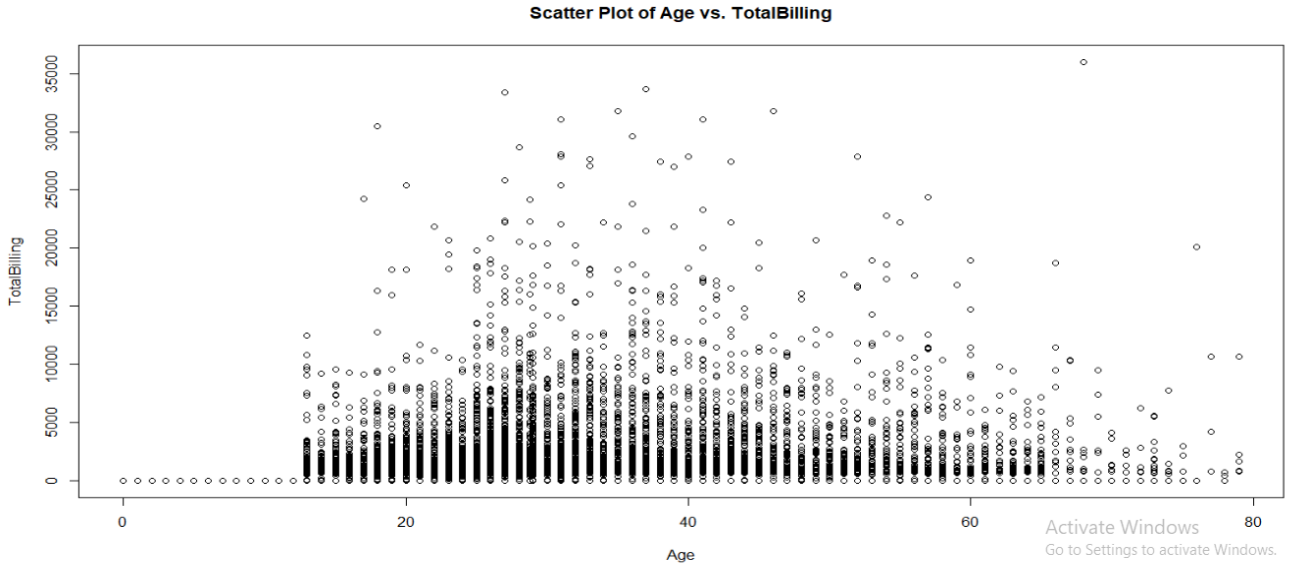
Correlation Analysis

Correlation between Age and TotalBilling to see if there's any relationship between age and spending

- **Correlation Coefficient:**

The correlation coefficient of 0.177767 suggests a weak positive relationship between Age and TotalBilling, indicating a slight tendency for higher spending as passengers get older. However, further investigation and analysis may be needed to fully understand the factors driving spending behavior among passengers.

- **Scatter Plot:**



The denser concentration of data points in the age range of twenty to sixty suggests that this demographic group comprises a significant portion of passengers and exhibits varied spending behaviors during their journey aboard the Spaceship Titanic. Further analysis can help uncover the underlying factors driving these spending patterns and inform strategies for catering to passengers' needs and preferences.

Group Comparisons

Comparison of the average TotalBilling between passengers in CryoSleep and those not in CryoSleep

The average TotalBilling for passengers in CryoSleep is **0**.

The average TotalBilling for passengers not in CryoSleep is **$2215.86**.

This indicates that passengers who are not in CryoSleep have a significantly higher average TotalBilling compared to passengers in CryoSleep. It's essential to note that the average TotalBilling for passengers in CryoSleep being 0 could be due to specific reasons, such as passengers in CryoSleep not being actively engaged in spending activities or not having access to amenities during their suspended animation.

This, further, suggests a potential relationship between CryoSleep status and passenger spending behavior, with passengers not in CryoSleep contributing more to the total revenue generated from onboard amenities. Further exploration and analysis could uncover additional insights into the factors driving these spending patterns and inform strategies for catering to passengers' needs and preferences.

Comparison of the average Age of VIP passengers and non-VIP passengers

Mean age for VIP passengers: **36.61** years

Mean age for non-VIP passengers: **28.60** years

This indicate that, on average, VIP passengers are older than non-VIP passengers. There is a significant difference in the average age between VIP passengers and non-VIP passengers. The average age difference between VIP and non-VIP passengers is approximately 8 years. This information can be valuable for understanding the demographics of passengers who opt for VIP services compared to those who do not.