**Introduction:**

Air travel is often underestimated, being solely perceived as a prelude and postlude to the common vacation experience. However, its significance increases vastly for those who embark on long international journeys. In this case, it is an integral part of the entire journey. Because of this, the quality of the experience matters substantially, with many aspects contributing to it, such as seat comfort, cabin staff service, ground service, value for money, and even food and beverage ratings.

Considering the undeniable impact of air travel on the overall vacation experience, my project aims to investigate how much these aspects affect the overall rating of the experience. Through the lens of newly acquired statistical knowledge, I pose the question: **How do seat comfort, cabin staff service, ground service, value for money, and food and beverage ratings collectively influence the overall rating of an airline?**

This investigation is more than a statistical conquest; it holds the key to understanding what makes a plane, which may seem mundane from the exterior, memorable. Understanding the nuances of passenger satisfaction allows us to draw conclusions that improve not only the experiences of travelers but also of airlines themselves. The quality of air travel can shape the story of our travels and craft a quality foundation for our vacations.

Another question that I wanted to answer is about the **quality of the data**. **Do incomplete entries with missing data affect the accuracy of a model?** I answered this by creating two scenarios, which I will explore more in the 'Data Description' section.

**Data Description:**

Focusing on my goal to regress the overall rating on various individual ratings, I identified two columns with significant data missing: Inflight Entertainment and Wifi and Connectivity. These factors are potential predictors that I could have used to help answer my general question, but with so much missing, I would have had an extremely small dataset to work with. This would result in a less accurate model, so **I unfortunately had to remove the Inflight Entertainment and Wifi and Connectivity columns**. I kept the filters on, which filtered out all the rows containing at least one blank value, and then copied that dataset to a new CSV file. I undid the filters to keep the missing values, then held that dataset as well.

```
##                  Column MissingCount
## 1         OverallRating            5
## 2           SeatComfort          116
## 3     CabinStaffService          127
## 4         GroundService          845
## 5         ValueForMoney            0
## 6        Food&Beverages          374
## 7 InflightEntertainment         1088
## 8     Wifi&Connectivity         3092
```

The dataset used to answer this question was web scraped from https://www.airlinequality.com/, compiled by Anshul Chaudhary and Muskan Risinghani, and hosted on Kaggle. My investigation began with a close analysis of the raw data in Excel, which revealed a multitude of missing values. To quantify this, I used the filter function in Excel, and created a table that revealed the amount of missing values per column.

**Note: I omitted the columns that are irrelevant of my general idea of regressing the overall rating on various ratings.**

**I wanted both versions of the datasets so that I had a fully clean version and a partially clean version.** I added both datasets to a folder in which I would have my R file so that it could be read in both datasets. So, my first scenario would be creating a model with the fully cleaned data, and my second scenario would be building a model with the data where there are missing values.

Next, I moved on to R, where I further cleaned the datasets. After reading them in, **I utilized the c (combine) function to remove the columns that weren't potential predictors or the overall**

**rating**. Here is what the first few entries of the dataset look like for both of the sets. These tables show the values for Overall Rating, Seat Comfort, Cabin Staff Service, Ground Service, Value for Money, and Food and Beverages. Observe some of the missing values in the Food and Beverages section.

```
##   OverallRating SeatComfort CabinStaffService GroundService ValueForMoney
## 1             3           2                 3             1             2
## 2             8           3                 3             4             3
## 3             1           1                 1             1             1
## 4             1           1                 1             1             1
## 5             8           5                 5             4             4
##   Food.Beverages
## 1              1
## 2              4
## 3              1
## 4              1
## 5              4
```

```
##   OverallRating SeatComfort CabinStaffService GroundService ValueForMoney
## 1             1           1                 1             1             1
## 2             3           2                 3             1             2
## 3             8           3                 3             4             3
## 4             1           3                 3             1             1
## 5             1           1                 1             1             1
##   Food.Beverages
## 1
## 2              1
## 3              4
## 4
## 5              1
```

With Overall Rating as my dependent variable, my regression analysis focuses on the aforementioned potential predictors: Seat Comfort, Cabin Staff Service, Ground Service, Value for Money, and Food and beverages. I aim to quantitatively evaluate each variable's impact on passengers' overall satisfaction. Each predictor variable's regression coefficient offers information about the strength and magnitude of its influence, allowing for a more detailed understanding of the factors that most strongly influence changes in the Overall Rating. It is also important to examine the effects of data with missing values on the model's accuracy to determine if cleaner data results in more reliable regression results. So, I will refer to the dataset with the omitted missing values as the **filtered data** and the other dataset as the **unfiltered data**.

Now, onto the subsets. I chose to take both sets of data and create two subsets each.

Before filtering there were 3701 entries in the dataset. After filtering for missing values there are 2537 entries in the dataset. **For my filtered data, I took a random 15% of the data (381 entries) to use as test values for my first model, which I will use the remaining 85% of the data (2156 entries) to build**. For my unfiltered data, I will use the same number of values as I did for the test and training values, so it will be easier to calculate the mean squared error. This will be used to compare the efficacy of both models.

**Methods:**

As stated previously, my approach to answering my second question was to have two datasets to create two models - one with data that filtered out the entries with blank values and one with empty values. By creating two different models using the airline dataset, I can examine the effects and potential consequences of the blank values. I can evaluate how the inclusion or exclusion of data points affects the accuracy, reliability, and generalization of the models. I can get a sense of how well the model can predict the overall rating

of reviews when missing values are excluded. By comparing the performance of both models, I can show insights into how well the model with incomplete data can predict the overall rating and whether it is strong enough to manage incomplete information.

My first approach to my regression analysis was implementing the Stepwise Regression algorithm to achieve the best of both worlds from the Forwards Selection and Backwards Elimination algorithms. When I was in the initial stage of fitting each predictor into its own model, one at a time, I used the linear regression summary to determine the p-values. This was done so I could compare the p-values to each other and decide which predictors to keep. However, all ten predictors boasted p-values below 2.2e-16. They were so low that R displayed them under this range instead of providing an exact value. This made it difficult to determine which predictors were more significant than others, so I ultimately chose not to follow the rest of the Stepwise Regression process. I kept all five predictors for both models, thus not fitting any reduced models.

Each predictor showed an incredibly low p-value, far below the standard threshold of 0.05 confidence. This was my reasoning for including every single predictor in the final models. Due to the high degree of statistical significance for all of the predictors, I realized that keeping all of the predictors in would fully reflect each individual rating on the overall rating. This was most definitely an exceptional situation. I imagine that this does not happen very often. However, one would generally get different p-values for each predictor and choose the predictors accordingly, depending on which selection algorithm they decide to go with.

**Results:**

Here are the summaries of both of the linear regression models:

```
##
## Call:
## lm(formula = training_set$OverallRating ~ training_set$SeatComfort +
##     training_set$CabinStaffService + training_set$GroundService +
##     training_set$ValueForMoney + training_set$Food.Beverages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9498 -0.6423  0.0903  0.7044  5.2032
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.81296    0.07337 -24.710   <2e-16 ***
## training_set$SeatComfort        0.34038    0.02785  12.223   <2e-16 ***
## training_set$CabinStaffService  0.24697    0.02685   9.198   <2e-16 ***
## training_set$GroundService      0.42377    0.02466  17.187   <2e-16 ***
## training_set$ValueForMoney      0.97993    0.03162  30.993   <2e-16 ***
## training_set$Food.Beverages     0.35348    0.02979  11.866   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.234 on 2150 degrees of freedom
## Multiple R-squared:  0.8417, Adjusted R-squared:  0.8413
## F-statistic:  2286 on 5 and 2150 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = OverallRating ~ SeatComfort + CabinStaffService +
##     GroundService + ValueForMoney + Food.Beverages, data = uf_training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.8613 -0.7105  0.1080  0.7155  5.1482
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.78805    0.08600 -20.791   <2e-16 ***
## SeatComfort        0.32270    0.03380   9.546   <2e-16 ***
## CabinStaffService  0.25696    0.03080   8.343   <2e-16 ***
## GroundService      0.43105    0.02849  15.129   <2e-16 ***
## ValueForMoney      0.94265    0.03730  25.274   <2e-16 ***
## Food.Beverages     0.38264    0.03508  10.907   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 1456 degrees of freedom
##   (694 observations deleted due to missingness)
## Multiple R-squared:  0.8491, Adjusted R-squared:  0.8486
## F-statistic:  1638 on 5 and 1456 DF,  p-value: < 2.2e-16
```

When examining these linear regression summaries, a plethora of insights emerge from the models. The filtered data model yields a highly significant and robust model, which can be told from the B values and adjusted R^2 value. The B value coefficients show a strong positive correlation, especially with **B4 (Value for Money) at 0.97993**. This essentially means that when the Value of Money rating increases by 1 unit, the overall rating increases by 0.979333, which is almost 1 unit. The **adjusted R^2 value of 0.8413 signifies that the model is statistically significant**, and the extremely low p-value for the model (less than 2.2e-16) also shows the null hypothesis that B1, B2, B3, B4, and B5 = 0.

In comparison, the unfiltered data model demonstrates statistical significance, even though the coefficient values are different. For example, the B4 value (Value for Money) is 0.94265, which doesn't equal the B4 value for the filtered data model. This slight difference shows a subtle impact of uncleaned data on the model's parameters. The cleaned data shows a higher correlation between the Value for Money and the Overall Rating. **The adjusted R^2 value is also different, with the unfiltered data model producing a value of 0.84846**. While this is still very close to 1, this may lead some to question the accuracy of this model since many may think that the unfiltered data model would result in a less accurate model.

These equations estimate the overall ratings based on the individual service factors. Both models reject the null hypothesis of the coefficients being 0, thus emphasizing the significance of the predictors in determining the overall ratings. Notably, the unfiltered model regularly shows lower coefficients for several variables, including food and beverages, ground service, cabin personnel service, and seat comfort, indicating the possible effects of missing data on these variables.

Even though these results offer substantial insights, there are definitely limitations that occur due to missing data in the filtered model. This results in challenges posed to the model's effectiveness and can limit the reliability of certain coefficient estimates.

After analyzing the coefficient values, p-values, and adjusted R^2 values, I calculated the Mean Squared Error (MSE) for both filtered and unfiltered data models. I chose this because I wanted a more robust method of determining which data models were more accurate. The Mean Squared Error essentially measures the average squared differences from the residuals of the models. It is obtained by summing the squared residuals and dividing that by the degrees of freedom for residuals.

Here are the MSE values from my data:

```
## [1] "MSE for Filtered Data: 1.52237941722175"
```

```
## [1] "MSE for Unfiltered Data: 1.42124157527593"
```

**The lower the MSE value, the better the performance of the model**. So, comparing the MSEs between the filtered and unfiltered data models can provide insights into the impact of missing data on the model's accuracy. The MSE for Unfiltered Data (approximately 1.42124) ended up being smaller than the MSE for the Filtered Data (about 1.52237). **This surprised me since I thought the presence of missing data would decrease the model's accuracy**.

**Conclusion:** In conclusion, the analysis of the predictor variables influencing the overall rating of an airline revealed interesting insights. The filtered data model, which excluded entries with missing values, produced a statistically significant model. The coefficients, particularly the one for Value for Money (B4), demonstrated a strong positive correlation with the overall rating, implying that when the value of money rating increases by 1 unit, the overall rating increases by 0.979333, almost 1 unit. The high adjusted R^2 value of 0.8413 also attests to the model's reliability.

On the other hand, the unfiltered data that incorporated entries with missing values also resulted in a statistically significant model with a high R^2 value but with slightly different coefficient values. The Value for Money coefficient was smaller than this coefficient in the filtered data model, suggesting a nuanced impact of missing data on this variable's correlation with the overall rating.

While both models reject the null hypothesis of coefficients being 0, the unfiltered model is more dependable, exhibiting a smaller Mean Squared Error value. **This challenges the widely accepted belief that missing data compromises the accuracy of a model**. However, my study is not without limitations. The exclusion of data entries from the filtered model may influence the reliability of certain coefficient estimates.

Expanding the dataset, specifically in the reviews for Inflight Entertainment and Wifi & Connectivity, can result in a more robust model, especially if these other individual reviews can be considered for the model. This results in a more powerful model and a more comprehensive understanding of how different factors of the airline experience impact the overall rating.