**Assignment-based Subjective Questions**

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- *Months during Summer/Fall and Winter contributed more bookings compared to Spring. This increased significantly during 2019 compared to 2018, indicating a good growth of business.*
- *Working day or not didn't matter that much.*
- *Holidays lead to a lower number of bookings compared to other days. Maybe people already had prior plans for their holidays and didn't prefer being outside on a bike.*
- *Increasing trend observed from Jan till Sep, then the booking count drops from Oct.*
- *Mostly clear weather contributed higher bookings among weather situations. With snow/light rain contributing the least.*
- *Summer, Fall and Winter seasons attracted more people compared to Spring.*

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

*We just need n-1 variables for encoding n states, not using this will leave an additional column leading to more features and increased correlation between features. For e.g.*

*Encoding 4 seasons requires only 3 dummy variables. If all of them are not set, it's obvious that this record maps to the one season that's left out.*

*Therefore, setting drop_first as true in get_dummies, drops this additional unneeded column.*

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

*Highest correlation is with temp/atemp variables.*

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- *No multicollinear terms, VIF less than 5 for all features. (Multicollinearity)*
- *Normally distributed error terms when plotted. (Normal Distribution)*
- *When plotted, Actual values and predicted values show a common linear pattern. (Linearity)*
- *No observable pattern in residuals when plotted. (Independent terms)*
- *No change in variance across error terms when plotted (homoscedasticity)*

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

*Features driving up demand –*

- *Temp (Temp: 4103.2348)*
- *Year (Yr: 2028.9952)*
- *Clear weather (wthr_mostly_clear: 676.2652)*
- *September (mnth_Sep: 539.6480)*

*Features driving down demand –*

- *Light snow or rain (wthr_snowrain_light: -1897.6647)*
- *Spring Season (season_Spring: -987.1736)*
- *July (mnth_July: -589.1049)*

*So, from a holistic standpoint the top 3 features that explain the variations (+ve or -ve) are Temperature (+ve), Year (+ve) and Light snow or rain (-ve).*

---

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a statistical approach for establishing a relation between a dependent variable (often referred as `y`) and one or more independent or predictor variables (denoted by $x_1$, $x_2$ ..). This can be used to find a close line or plane that best fits the various data points. This is a regression line (when there is just 1 predictor variable) or hyperplane (multiple predictors).

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ..$$

$\beta_0$ is the intercept and $\beta_1$, $\beta_{2..}$ represent the coefficients for each predictor variables.

Y is the dependent variable and $x_1$ , $x_2$ .. represent the predictors.

This is best suited for situations which exhibit a linear relationship. There can be a positive or negative linear relationship between the dependent variable and the predictors.

Simple linear regression tries to identify the coefficients for the following equation by using just 1 predictor variable to explain Y –

$$Y = \beta_0 + \beta_1 * x_1$$

The coefficients are computed in such a way that it best fits all the data points. This is done via minimizing the differences between predicted Y and actual Y, typically via a cost function. For e.g., Ordinary Least Squares, Differentiation or Gradient Descent method.

Whereas Multiple linear regression uses additional variables instead of just 1 to better fit the line or in this case the hyperplane. The goal is to compute coefficients for the hyperplane to fit the data points provided, this is done via the least squares method or other suitable methods.

**Some assumptions of Linear Regression:**

- There should be some form of linear relation between Y and $x_1$ ,$x_2$, $x_3$, etc. Otherwise, there is no point in fitting a linear model between them.
- Error terms are normally distributed with mean at 0.
- Error terms are independent of each other and show no pattern when plotted.
- Error terms have constant variance, variance should not follow any patterns.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics (such as mean/median/std dev. etc.), but there are peculiarities that fool the regression model once you plot each data set. This dataset was developed by statistician Francis Anscombe.

The data sets (shown below) have very different distributions, so they look completely different from one another when you visualize the data on scatter plots.

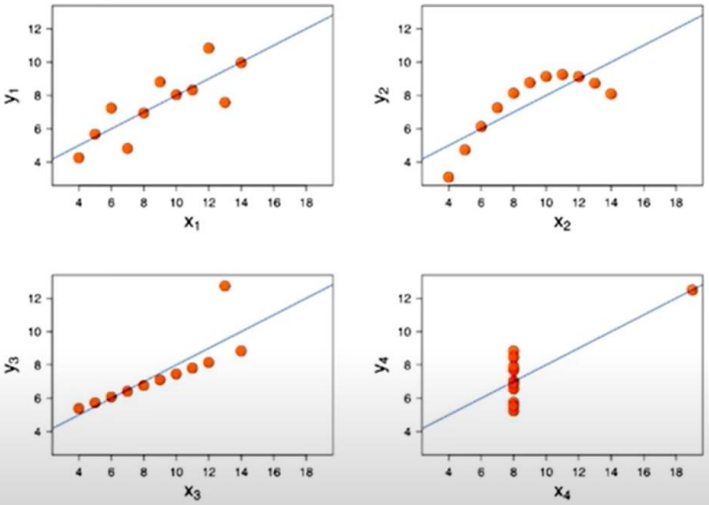
However, statistics on Anscombe's quartet will show that the data has identical features –

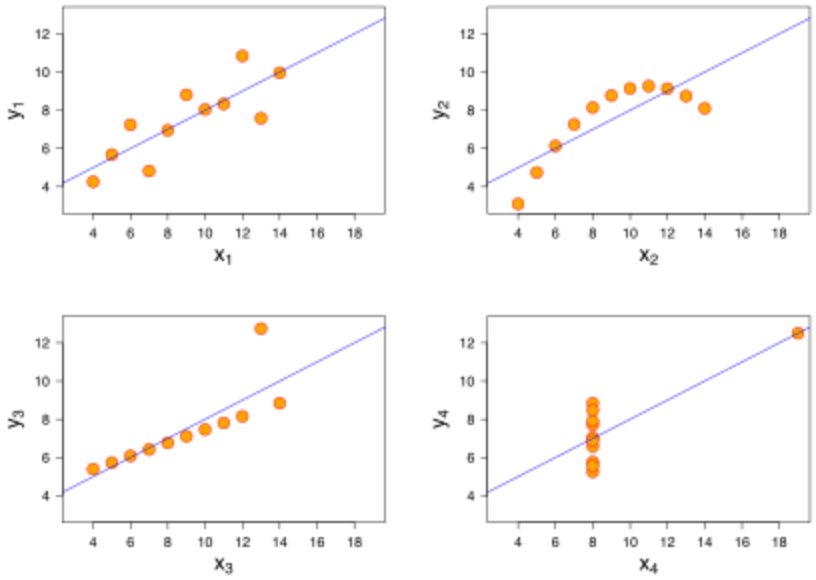
For all 4 datasets:

- Mean of x is 9.
- Sample variance of x is 11.
- Mean of y is 7.50
- Sample variance of y is 4.125
- Correlation between x & y is 0.816

When a regression line is plotted for these 4 datasets, we will arrive at the same line but visually the fit will look drastically different.

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |




Data Set 1: fits the linear regression model well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

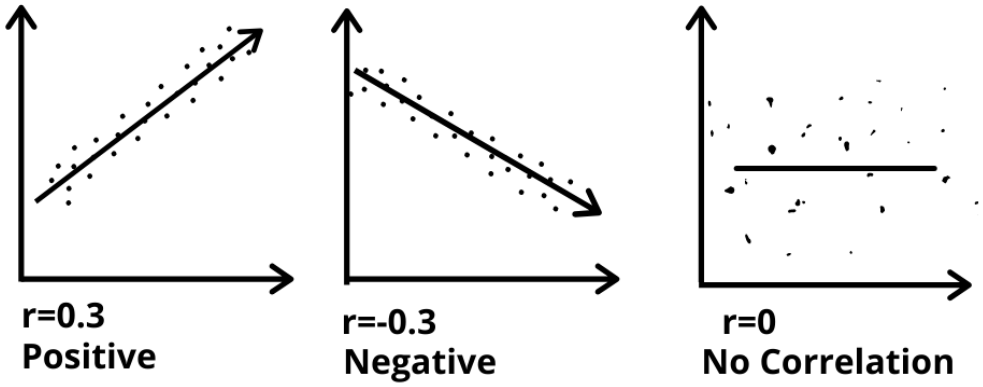
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R? (3 marks)

Pearson's R or Pearson Correlation Coefficient is a measure of linear correlation between 2 sets of data. It can range from -1 to +1. A value of 0 indicates no correlation whereas -1 indicates perfect negative correlation and +1 perfect positive correlation, with intermediate values indicating strength.

- $0 < R \leq +1$ => As one variable increases, so does the other variable
- $-1 \leq R < 0$ => As one variable increases, the other decreases.
- $R = 0$ => No association between the 2 variables.

**Example plots –**




4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a procedure to bring all independent variables on a common ground. If independent variables are all on varying units/scales, the model coefficients will be off range/weird and it becomes different to interpret the relationship. Scaling re-evaluates the features without affecting other parameters of the model such as - t-statistic, F-statistic, p-values, R-square etc.

Scaling does 2 things that are beneficial for us –

- Improves ease of interpretation
- Faster convergence for gradient descent methods.

Comparison between Normalized and Standardized scaling techniques –

| **Aspect** | **Normalized Scaling (Min-Max)** | **Standardized Scaling** |
|---|---|---|
| **Computation?** | $X_{\text{Normalized}} = \frac{X - X_{min}}{X_{max} - X_{min}}$ | $X_{\text{Standardized}} = \frac{X - X_{mean}}{X_{std}}$ |
| **Values after scaling** | Within [0,1] | Centered around 0, Std. Dev is 1. |
| **Effect on data** | Mean not preserved. | Data centered around 0. |
| **Effect on Std. Dev** | Std. Dev. not preserved | Scales to have Std. Dev as 1. |
| **Presence of outliers** | Distorted due to outliers as max-min range is inflated. | Not that affected by outliers as we are reliant on mean & SD instead of actual range. |
| **Available as?** | MinmaxScaler in ScikitLearn | StandardScaler in ScikitLearn |

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This can happen when the R-squared ($R_i^2$) is 1, implying that the variable can be perfectly explained by other variables. It's a case of perfect multicollinearity, leading to $1/(1-R_i^2)$ to become ∞.

Once the variable that's causing this perfect collinearity is dropped, VIF value should drop down from infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

A **Q–Q plot** is a plot of the quantiles of two distributions against each other. We plot the sample quantiles against theoretical quantiles (such as that of a normal distribution).

It can also help in verifying that the test and train datasets are from the same population. When the datasets are plotted, they should fall relatively on the same line then they are from the same population. If there is deviation, it indicates they are different populations or they are scaled differently.

It can also help in verifying one of the assumptions in Linear Regression – normality of residual values. After creating the model and obtaining residual values, we can plot the residual values on a Q-Q plot. If the terms lie on an approx. 45 deg straight line, then residuals are normally distributed. If it deviates heavily from the straight line, then this assumption might not hold true implying there might be accuracy & reliability issues on this model.