# 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)**

   **Answer**:

   I've used the boxplot and bar plot to analyse categorical columns. The following are a few conclusions we may get from the visualisation –

   - The fall season appears to have gotten more bookings. And, from 2018 to 2019, the number of bookings has risen dramatically in each season.
   - The majority of reservations were made in the months of May, June, July, August, September, and October. As the year progressed, the trend grew until the middle of the year, when it began to decline as we approached the conclusion of the year.
   - Clear skies attracted more bookings, as one might expect.
   - Bookings are higher on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
   - When it is not a holiday, the amount of bookings appears to be lower, which makes sense
   - Bookings appeared to be nearly equal on working and non-working days.
   - 2019 saw an increase in bookings over the previous year, indicating positive business growth.

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

   **Answer:**

   It's important to use drop first = True since it reduces the extra column formed during fake variable construction. As a result, the correlations between dummy variables are reduced.

   Drop first: bool, default False, which indicates whether to remove the first level from k category levels to generate k-1 dummies.

   Let's imagine we have three different types of entries in the Categorical column and want to make a dummy variable for it. If one of the variables isn't A or B, the obvious answer is C. As a result, we don't require the third variable to find C.

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

   **Answer:**

   The variable 'temp' has the strongest link to the target variable.

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

   **Answer:**

   1. Normality of error terms
      - Error words should be delivered normally.
   2. Multicollinearity check
      - Multicollinearity between variables should be negligible**.**

3. Linear relationship validation
    - The presence of linearity among variables should be obvious.
4. Homoscedasticity
    - In residual values, there should be no discernible pattern.
5. Independence of residuals
    - No autocorrelation exists.

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

**Answer:**

The top three characteristics that significantly contribute to explaining the demand for shared bikes are shown below –

- Winter
- Temperature
- September

# General Subjective Questions

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

**Answer:**

The statistical model that analyses the linear connection between a dependent variable and a set of independent variables is known as linear regression. The term "linear connection" refers to the fact that when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes as well (increase or decrease).

The following equation can be used to illustrate the relationship mathematically:

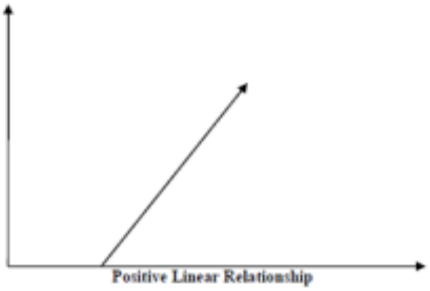
$Y = mX + c$

The dependent variable we're seeking to forecast is Y.

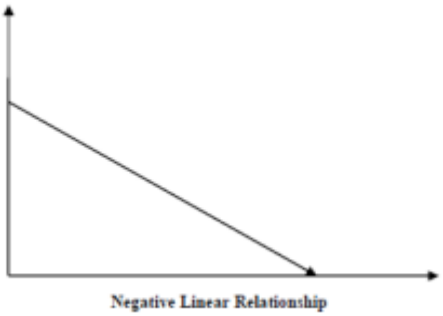
We are making predictions using the independent variable X.

m is the slope of the regression line, which reflects the effect of X on Y, and c is the Y-intercept, which is a constant. If X equals 0, Y equals c.

1. Positive Linear Relationship:
    - If both the independent and dependent variables rise, the relationship is said to be positive. The following graph can help you understand it.




2. Negative Linear relationship:
    - If the independent variable rises while the dependent variable falls, the connection is said to be positive. The following graph can help you understand it.




There are two types of linear regression:

- simple linear regression and
- multiple linear regression.

**Assumptions**:

The Linear Regression model makes the following assumptions about the dataset:

1. Multi-collinearity —
    - The linear regression model implies that the data has little to no multi-collinearity. Multi-collinearity is defined as when independent variables or features are dependent on one another.
2. Auto-correlation —
    - Another assumption made by the linear regression model is that the data has little or no auto-correlation. Auto-correlation arises when residual errors are dependent on one another.
3. Variables and their relationships—
    - The linear regression model implies that the relationship between the response and the feature variables is linear.
4. Normality of error terms—
    - Error terms should be distributed naturally.

5. Homoscedasticity—
    - There should be no evident pattern in residual values due to homoscedasticity.

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

**Answer:**

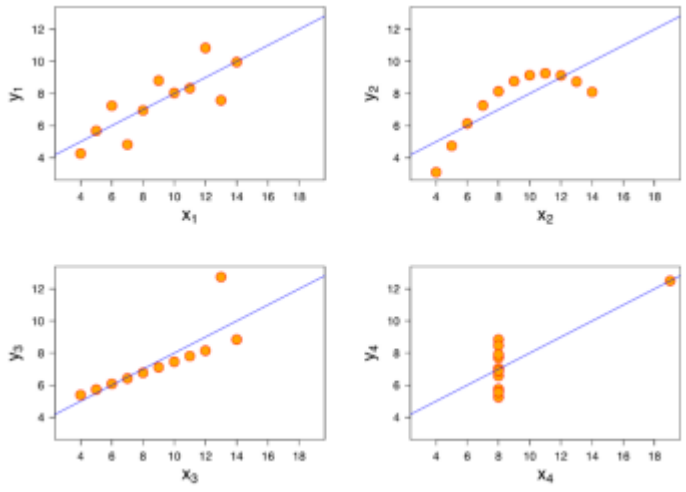
Francis Anscombe, a statistician, invented Anscombe's Quartet. It is divided into four datasets, each with eleven (x, y) pairings. The most important thing to keep in mind regarding these datasets is that they all use the same descriptive statistics. When things are graphed, however, they change totally, and I emphasise fully. Regardless of the identical summary data, each graph reveals a different narrative.

| | I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|---|
| | x | y | x | y | x | y | x | y |
| | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| | 11 | 8,33 | 11 | 9,26 | 11 | 7,81 | 8 | 8,47 |
| | 14 | 9,96 | 14 | 8,1 | 14 | 8,84 | 8 | 7,04 |
| | 6 | 7,24 | 6 | 6,13 | 6 | 6,08 | 8 | 5,25 |
| | 4 | 4,26 | 4 | 3,1 | 4 | 5,39 | 19 | 12,5 |
| | 12 | 10,84 | 12 | 9,13 | 12 | 8,15 | 8 | 5,56 |
| | 7 | 4,82 | 7 | 7,26 | 7 | 6,42 | 8 | 7,91 |
| | 5 | 5,68 | 5 | 4,74 | 5 | 5,73 | 8 | 6,89 |
| SUM | 99,00 | 82,51 | 99,00 | 82,51 | 99,00 | 82,50 | 99,00 | 82,51 |
| AVG | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 | 9,00 | 7,50 |
| STDEV | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 | 3,32 | 2,03 |

The averages and variances for x and y across the groups were identical, according to the summary statistics:

- For each dataset, the mean of x is 9 and the mean of y is 7.50.
- For each dataset, the variance of x is 11 and the variance of y is 4.13.
- For each dataset, the correlation coefficient (how strong a relationship exists between two variables) is 0.816.

We can see that these four datasets have the same regression lines when plotted on an x/y coordinate plane, but each dataset tells a distinct story:




- Dataset I looks to feature linear models that are clean and well-fitting.

- Dataset II is not regularly distributed.
- In Dataset III, the distribution is linear, but an outlier throws off the estimated regression.
- Dataset IV demonstrates that a single outlier can result in a high correlation coefficient.

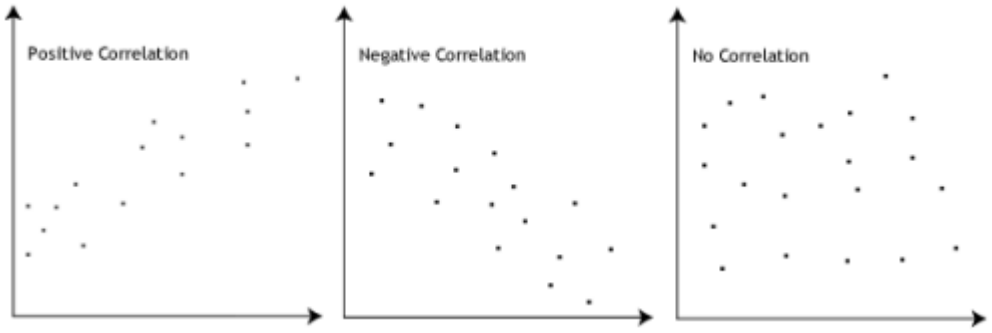
The importance of visualisation in data analysis is emphasised in this quartet. When you look at the data, you can see a lot of the structure and get a good image of the dataset.

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

**Answer:**

Pearson's r is a numerical representation of the strength of the linear relationship between variables. The correlation coefficient will be if the variables tend to go up and down together positive. If the variables tend to rise and fall in opposite directions, with one variable having low values. The correlation coefficient will be negative when one variable has high values.

The Pearson correlation coefficient, r, can be anything between +1 and -1. A value of 0 implies that the two variables have no relationship. A positive connection is defined as when the value of one variable grows, the value of the other variable increases as well. A negative relationship is indicated by a value less than 0; that is, as the value of one variable rises, the value of the other variable falls. The graphic below depicts this:




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

**Answer:**

Feature scaling is a strategy for putting the data's independent features into a set range. It is used to handle significantly changing magnitudes, values, or units during data pre-processing. If feature scaling is not done, a machine learning algorithm will assume larger values to be higher and smaller values to be lower, regardless of the unit of measurement.

For example, if an algorithm does not use the feature scaling approach, it may believe the value 3000 metres to be greater than 5 kilometres, which is not the case, and the algorithm will make incorrect predictions. So, to solve this problem, we employ Feature Scaling to bring all values to the same magnitude.

| S.NO. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |

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

**Answer:**

VIF = infinity if there is perfect correlation. The presence of a correlation between the variables is indicated by a high VIF value. If the VIF is 4, it signifies that multicollinearity has inflated the variance of the model coefficient by a factor of four.

When VIF is infinite, it indicates that two independent variables are perfectly correlated. We get R-squared (R2) =1 in the event of perfect correlation, which leads to 1/ (1-R2) infinite. To remedy this, we must remove one of the factors that is producing the perfect multicollinearity from the dataset.

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

**Answer:**

The quantile-quantile (q-q) plot is a graphical tool for detecting whether two data sets are from the same population.

The quantiles of the first dataset are plotted against the quantiles of the second dataset in a q-q graphic. The fraction (or percent) of points below a certain value is referred to as a quantile. That is, the 0.3 (or 30%) quantile is the number at which 30% of the data falls below and 70% falls above. Also plotted is a 45-degree reference line. The points should fall roughly along this reference line if the two sets originate from the same population with the same distribution. The larger the deviation from the reference line, the more evidence there is.

for the conclusion that the two data sets came from separate populations.

The significance of the Q-Q plot:

When two data samples are present, it is frequently desirable to determine if the assumption of a shared distribution is valid. If this is the case, estimators of location and scale can combine the

two data sets to generate estimates of the common position and scale. It is also beneficial to have some comprehension of the differences if two samples differ. Analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests can provide greater insight into the nature of the difference than the q-q plot.