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 :

   From Data visualization below our few things we can infer about the independent variables effect on dependent variable
   - Increase in bookings from Year 2018 to 2019 . Good progress in business
   - Months during the middle of the year have more bookings such as high bookings in months of May, Jun, July, Sept and Oct
   - Weather plays and import role in bookings and clear weather had most books
   - During holidays bookings are lower
   - Working day variable was not a very huge factor in determining the number of bookings

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

   We should use **drop_first=True** as it reduces the number of independent variables and helps reduce correlations across independent variables

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

   Temperature ('Temp') has the highest correlation with the target variable

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

   Below are the methods used to validate the Linear regressions assumptions
   - Error terms distribution : Error terms are distributed normally with 0 mean
   - VIF Check : Multicollinearity check and it was less the 4 for all Independent variables
   - Homoscedasticity : We did not have any visible patterns in residual values distribution
   - Linearity : we checked the linearity and had linear relationships

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 : Three main variables for explaining demand of shared bikes

   - Temperature
   - Season (Winter, Spring)
   - Month (Sep)

General Subjective Questions

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

   Answer :

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.
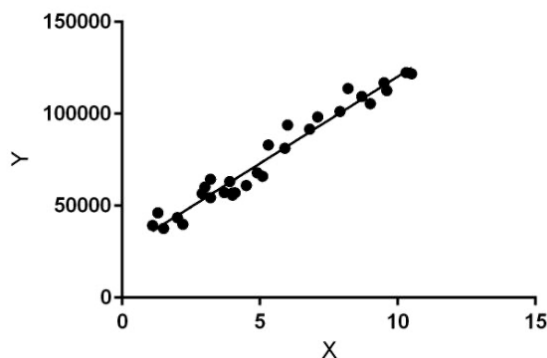
   Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

   In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

   Hypothesis function for Linear Regression :

   $$y = \theta_1 + \theta_2.x$$

   While training the model we are given :

   x: input training data (univariate – one input variable(parameter))

   y: labels to data (supervised learning)

   When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

   θ1: intercept

   θ2: coefficient of x

   Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x

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

Answer :

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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 summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:
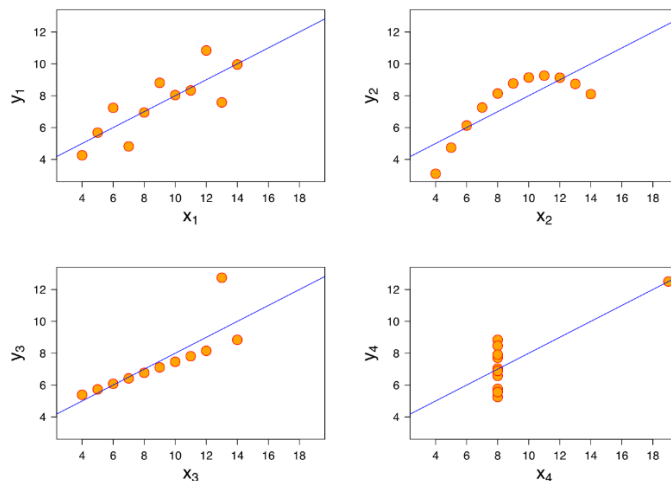
• Dataset I appears to have clean and well-fitting linear models.

• Dataset II is not distributed normally.

• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

   Answer :

   The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

   | Positive Correlation | No correlation | Negative |
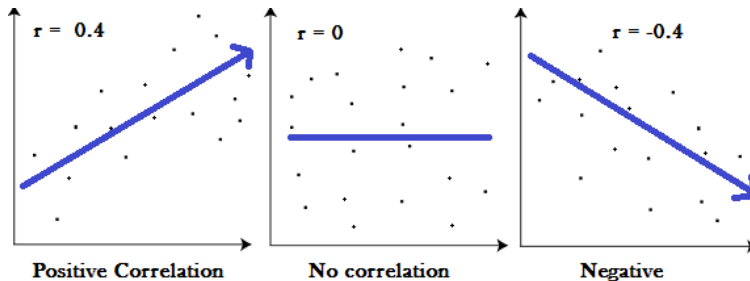
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** **(3 marks)**

   Answer :

   What : It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   Why : Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

   Difference :

   - **Normalized Scaling** : It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

   $$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   - **Standardized scaling** : Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

   $$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

    (3 marks)

    Answer :

    If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

    An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

    (3 marks)

    Answer. : Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.