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 the analysis done on categorical columns using the boxplot. Below are the few observations, we can infer –

- Summer/ Fall season seems to have more booking of bikes.
- This can be also seen in month graph too, most of the bookings has been done between June to October months.
- Clear weather has more booking, which can be understood.
- Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
- During holidays bike bookings are less compared to non-holidays.
- No much differentiation observed between working day and non-working day.
- 2019 have a greater number of bookings from the previous year, which shows good trend in terms of business.

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

Answer:

In pandas.get_dummies there is a parameter i.e. drop_first allows us to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels). Please note that drop_first = False meaning that the reference is not dropped and k dummies created out of k categorical levels! You set drop_first = True, then it will drop the reference column after encoding.
In case if Multiple Linear Regression. Keeping k dummies for k levels of a categorical variable creates redundancy of one level, which is here in separate column. This is not needed since one of the combinations will be uniquely representing this redundant column. Hence, it’s better to drop one of the columns and just have k-1 dummies(columns) to represent k levels.

This Overall approach reduces Multi-colinearity in the dataset, which is one of the prime assumption of Multiple Linear Regression.

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

Answer:

'atemp' variable 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:

No auto-correlation or independence

The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms of the time series data.

Durbin-Watson value of final model is 2.032, which signifies there is no autocorrelation.

No Multicollinearity

The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model. In essence, it is difficult to explain the relationship between the dependent and the independent variables.

VIF of final model has features which have values < 5 which no multicollinearity.

Homoscedasticity

Homoscedasticity means the residuals have constant variance at every level of x.

Scatter plot generated between residual vs fitted value shows no pattern, it means the residuals have constant variance (homoscedasticity).

Normal distribution of error terms:

The last assumption that needs to be checked for linear regression is the error terms' normal distribution. Error terms plot clearly shows they are normally distributed.

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:

Based on the final model, top 3 features contributing significantly towards explaining the demand of the shared bikes are **“atemp”, “Sep”(month),”winter”(season).**

General Subjective Questions

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

Answer:

Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

Linear regression is of the 2 types:

I. **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

   a. **Formula for the Simple Linear Regression:** Y=β0+β1X1 +ϵ

II. **Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a 'hyperplane' instead of a straight line.

Formula for the Multiple Linear Regression: Y=β0+β1X1+β2X2+…+βpXp+ϵ

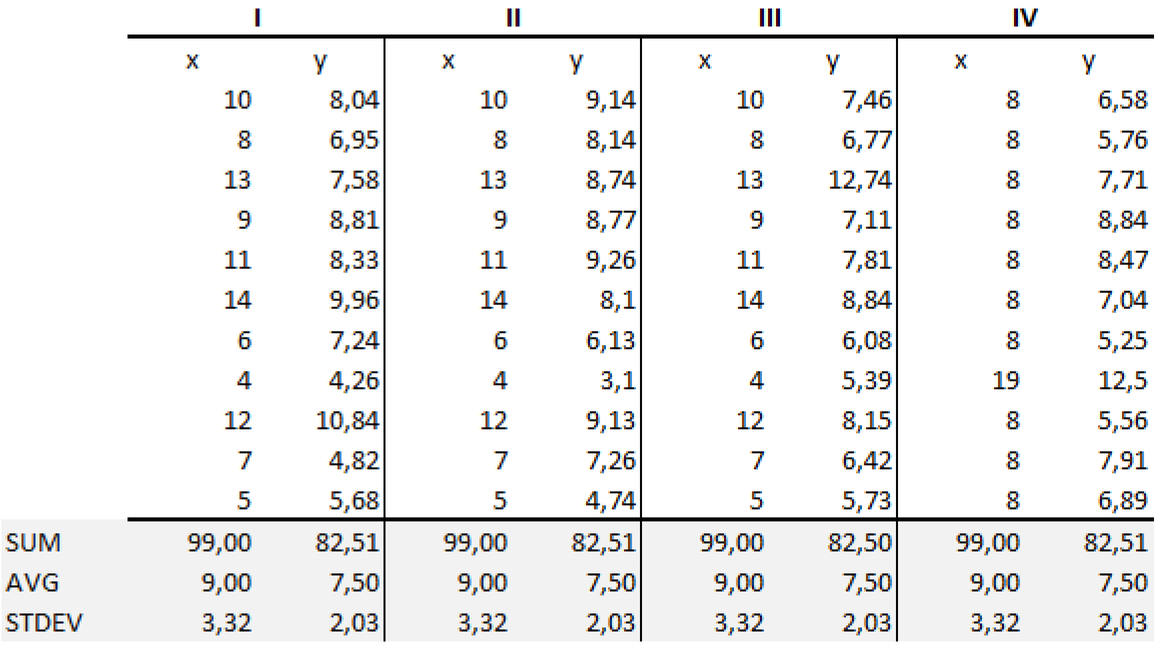
The equation of the best fit regression line $Y = \beta_o + \beta_1 X$ can be found by the following two methods:

- Differentiation.

- Gradient descent

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

Answer:
Anscombe's Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:



	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

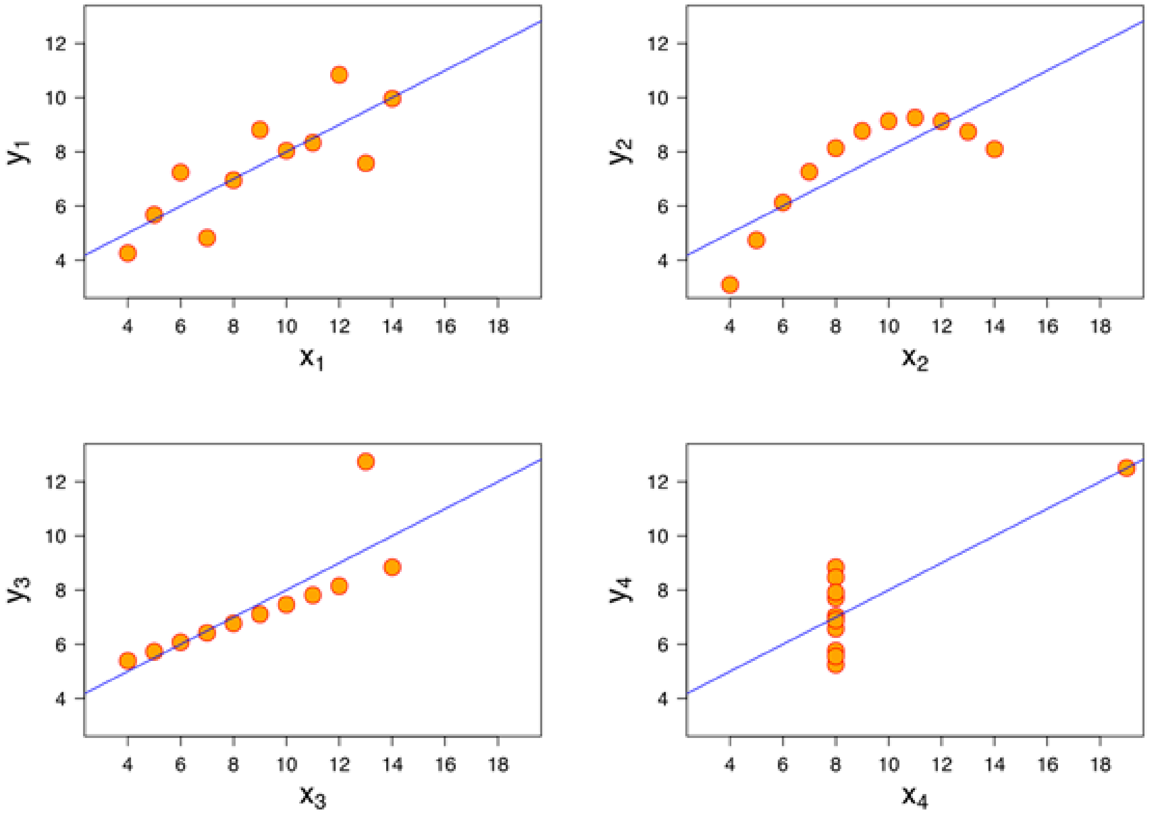
The summary statistics show that the means and the variances were identical for x and y across the groups:

· 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.

Additionally, Anscombe’s Quartet warns of the dangers of outliers in data sets.

3. What is Pearson’s R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

$r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

$r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

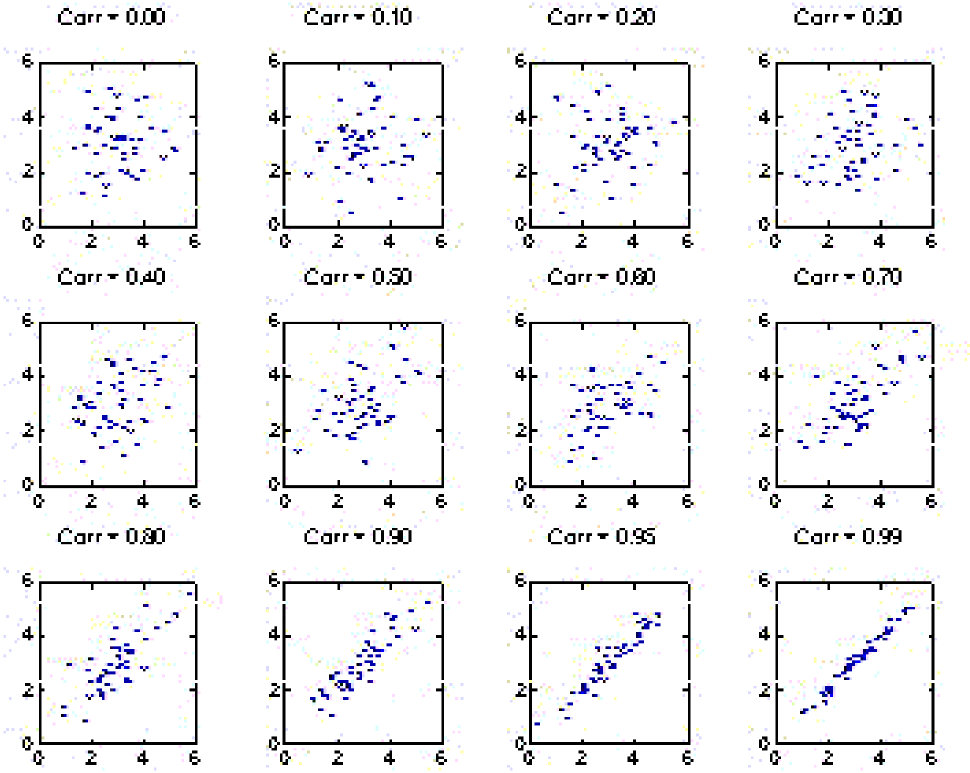
$r = 0$ means there is no linear association

$r > 0 < 5$ means there is a weak association

$r > 5 < 8$ means there is a moderate association

$r > 8$ means there is a strong association

The figure below shows some data sets and their correlation coefficients.



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

What is Feature Scaling?

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Why is scaling performed?

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.

By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

What is Normalization?

Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

What is Standardization?

Standardization is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that, in this case, the values are not restricted to a particular range.

Normalize vs Standardize:

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

It is a good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process. Also, the scaling of target values is generally not required.

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. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

Where, 'i' refers to the ith variable.

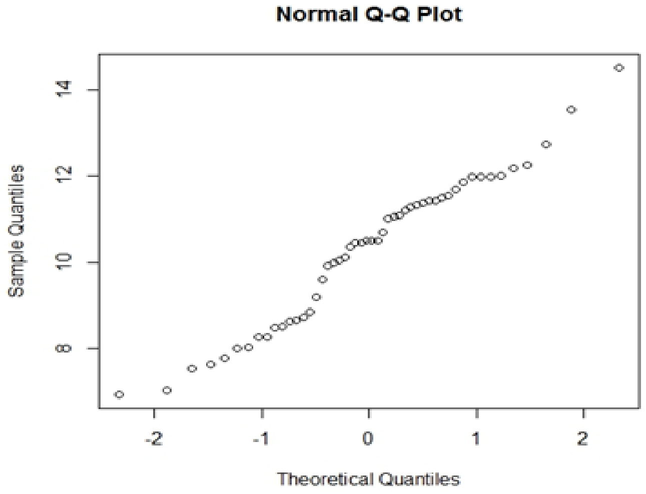
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

Answer:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. III. The q-q plot can provide more insight into the nature of the difference than analytical methods.