Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans-

- The categorical variables year , month, season, weather condition seems to have high correlation with the count (the dependent) variable.
- Demand for bikes is more in year 2019 than in 2018.
- Demand decreases as the weather goes from good to moderate to bad weather condition. The demand is significantly less for the weather condition 3 which has light rain+ thunderstorm+ scattered clouds and light snow.
- In case of seasons, demand is less for spring and winter season than for summer and fall season. This trend can also be seen in months, as Jan, Feb, March has spring in US and Oct, Nov, Dec are winter season (which have low demand) and the demand increases from April to September which is summer and fall season. We can conclude there is a relation between season and month column as well.
- Working day, holiday and weekday variable does not affect the dependent variable much (seems to have low correlation). The median (and the range) for the bike demand is little lower if it is a holiday.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans-

- It is important to use drop_first=True during dummy variable creation as it drops the first dummy column which reduces the multicollinearity. One of the dummy columns can always be explained by the rest therefore, using drop_first=True reduces cyclical dependency .
- Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
- Example- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one row has not C1 and not C2, then it is obviously C3. So we do not need 3rd variable to identify the C3.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

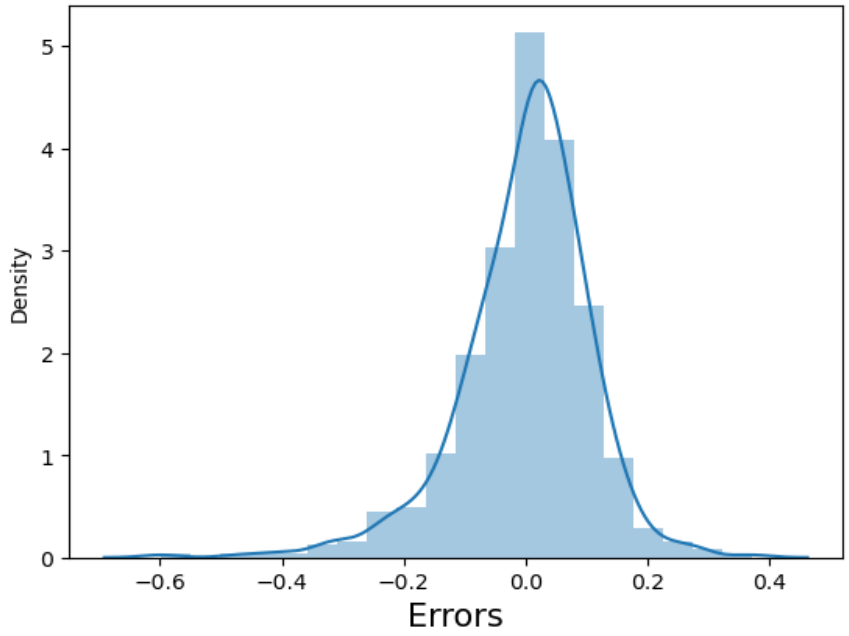
Ans-

The temp and atemp variable (i.e. temperature and feeling temperature) have the highest correlation with the dependent variable as seen from the pairplot.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

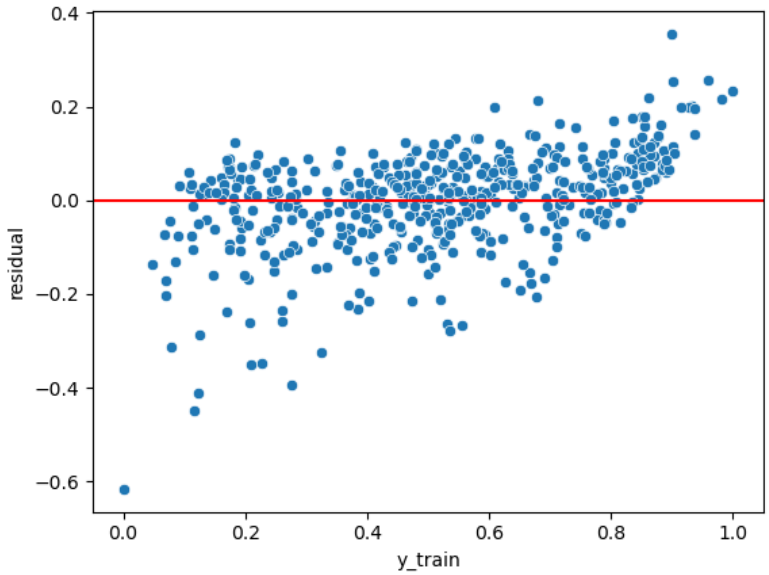
Ans- Assumptions are validated as follows-

- **Error terms are normally distributed with mean zero**- This has be justified by a distribution plot (shown below) for the error terms i.e. (actual-predicted) values.



The error terms are normally distributed with mean 0 in the plot.

- **Errors are independent of each other**.
  There is no visible pattern seen in the scatter plot below.
- **Homoscedasticity-** As can be seen from the plot below, errors are constant along the dependent variable.



- **There is no multicollinearity between the independent variables**. VIF <5 for all the predictors of the final model assures this.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans-

The **top three variables are temperature, year and windspeed**. All three have a very low p value (0.00 approx) and absolute value of coefficients of these three predictors are higher than others.

- Temperature - Coefficient of temperature indicates that a unit increase in temperature variable, will **increase** bike demand by **0.4599** values.
- Year - Coefficient of year indicates that a unit increase in year variable, will **increase** the bike demands by **0.2400** values.
- Windspeed - Coefficient of windspeed indicates that a unit increase in windspeed variable, will **decrease** the bike demand by **0.1694** values.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans-

Linear regression is a statistical model that analyses the linear relationship between a dependent variable and a single or a set of independent variables. Linear relationship between dependent and independent variable implies that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Linear regression can be divided into two major categories –

- **Simple Linear Regression**:

  Mathematical relation- **y=mX+c**,

where y is the dependent variable (which we want to predict), X is the independent variable also known as predictor, m denotes the slope of the regression line which represents the change in variable Y for one unit change in variable X , c is a constant, also known as the y-intercept. If X = 0, y will be equal to c.

- **Multiple Linear Regression:**

  Mathematical relation - **y=aX1+bX2+cX3+…. +nXn+ beta**

where y is the dependent variable (which we want to predict), X1, X2, … ,Xn are the independent variables (which we are using to make predictions) also known as predictors, a denotes the change in variable y for one unit change in variable X1, keeping all other independent variables (X2,X3,…Xn) constant, beta is a constant, it denotes the value of y if all the independent variables are 0.

ASSUMPTIONS:

The following are some assumptions that are made by a Linear Regression model –

- **No Multi-collinearity** – Linear regression model assumes that there is very little or no multi-collinearity in the data. Multi-collinearity occurs when the one or more independent variables are dependent on others. In other words, one independent variable can be described using other independent variable(s).
- **Error terms are independent of each other** – The error terms should not be dependent on one another
- **Linear relationship between the target variable** (dependent variable) **and the feature variables** (independent variables).
- **Normality of error terms**- Error terms are normally distributed with zero mean.
- **Homoscedasticity**- Error terms have constant variance.

Q2. Explain the Anscombe's quartet in detail.

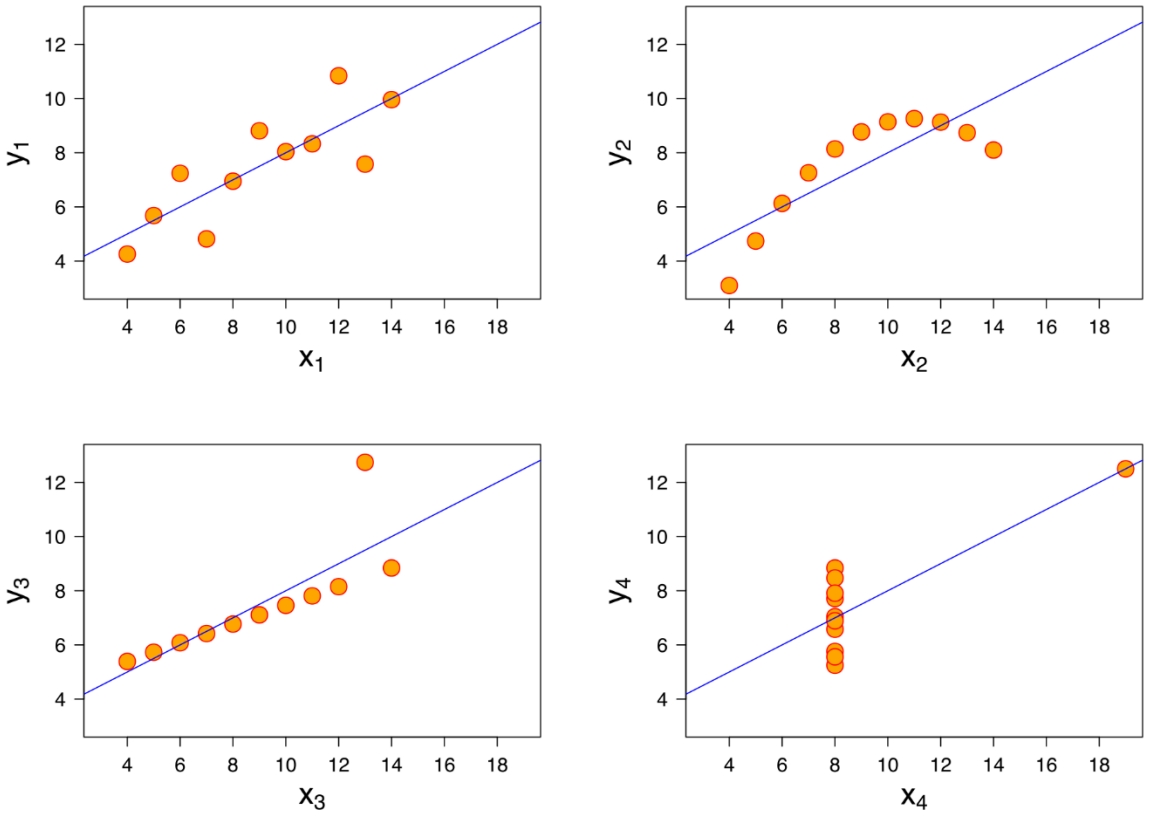
Ans-

Anscombe's quartet emphasizes on the importance **of visualization in Data Analysis**. It shows that looking at the data (plots) reveals a lot of the structure and gives a clear picture of the dataset. The summary statistics is not enough to get the entire insight.

Anscombe's quartet is a group of four datasets that have the same mean, standard deviation, and regression line, (descriptive statistics) but have different distributions. Each dataset consists of eleven (x,y) points. These were constructed in by a statistician Francis Anscombe, to demonstrate **both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.**

	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 in the figure above shows **that the means and the variances are identical for x and y across the four groups.**



From the plots of these four datasets, we can observe that although they **show the same regression lines but each dataset is telling a different story**:

- Dataset I have well-fitted linear models.

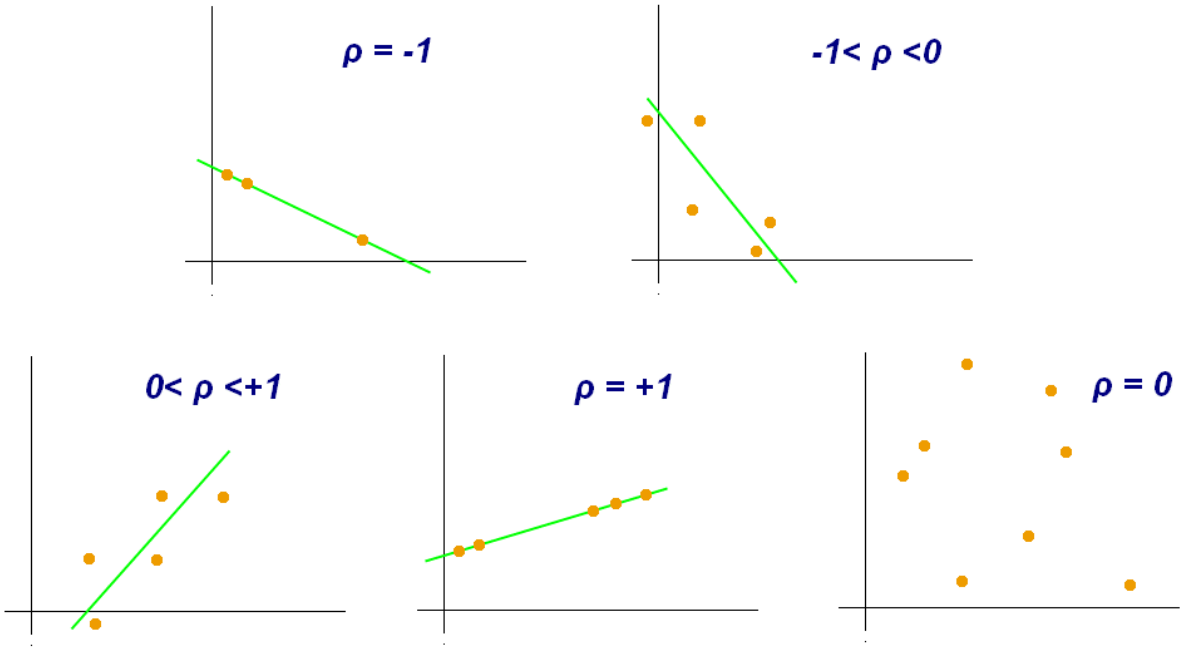
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the regression line is thrown off by an outlier.
- Dataset IV is completely different from these and shows that one outlier can produce a high correlation coefficient.

Q3 What is Pearson's R?

Ans-

The Pearson correlation **measures the strength of the linear relationship between two variables**. The **value of Pearson's correlation R lies between -1 and 1**, where,

- a value of -1 indicates there is a total negative linear correlation between two variables
- 0 means there is no correlation
- + 1 indicates, there exists a total positive correlation between two variables.



- A Pearson's correlation of 0 indicates that there is no association between the two variables.
- If the two variables tend to increase or decrease together, the correlation coefficient will be positive.
- If on increasing the value of one of the variable, the other variable decreases or vice versa, then the correlation coefficient will be negative.

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans-

Feature scaling is the process of normalising the range of features in a dataset.

Real-world datasets often contain features that varies in degrees of magnitude, range and units .Feature scaling is performed during the data pre-processing , **to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values regardless of the unit of the values.**

For example: If an algorithm does not uses feature scaling method, then it can consider the value 3000 g to be greater than 5 kg but that's actually not true and in this case, and this might give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes.

Machine learning algorithms rely on gradient descent to minimise their loss functions. Therefore, to **ensure that gradient descent converges more smoothly and quickly**, we need to scale our features so that they share a similar scale.

Majorly, there are two types of scaling-

- **Normalisation-**

1. Normalisation (or min-max scaling), is a scaling technique where the values in a column are shifted so that they lie between a fixed range of o and 1.
2. Minimum and maximum value of features are used for scaling.
3. It is really affected by outliers
4. It is used when features are of different scales.
5. MinMaxScaler in the Scikit-learn library is used for normalisation in Python.

- **Standardisation-**

1. Standardisation is a scaling technique where the values in a column are rescaled so that mean becomes 0 and variance becomes 1 i.e. they demonstrate the properties of a standard Gaussian distribution.
2. Mean and standard deviation is used for scaling. It is not bounded to a certain range.
3. It is much less affected by outliers.
4. It is used when we want to ensure zero mean and unit standard deviation.
5. StandardScaler in the Scikit-learn library is used for standardisation in Python.

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

Ans-

If VIF = infinity, then the variable (whose VIF=inf) is **perfectly and strongly correlated with another variables**. For example, if the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity. VIF is given by

VIF =1/ (1-R2)

In the case of perfect correlation, we will **get R-squared (R2) =1**, which will tend **VIF = 1/ (1-R2) to infinity.**

To solve this problem**, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.**

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

Ans-

Q-Q plot is also referred as quantile-quantile plot. **It is a plot between the quantiles of the two datasets.** Quantile means the fraction (or percent) of the points below the given value.

For example, 0.4 (or 40%) quantile is the point at which 40% percent of the data fall below and 60% fall above that value.

Use and Importance in linear regression-

Q-Q plot determines if the two datasets are coming from the populations have same distribution or not.

A 45-degree line known as reference line is plotted. If the two sets come from a population with the same distribution, the points should fall approximately along the reference line. The greater the distance of points from the reference line, we can conclude with a greater the evidence that the two data sets have come from populations with different distributions.

When there are two data samples, it is often desirable to know if the assumption of a same distribution for both is justified. The q-q plots helps here, to determine if the above mentioned assumption is met or not i.e. if the two data samples come from same distribution or not.

q-q plot helps in linear regression top check if the training and test dataset are from the same distributions or not. We might have training dataset and test dataset coming separately, so we can use q-q plot to confirm if both the datasets are coming from populations with same distribution or not.