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?**

Solution: I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

1. Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
2. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
3. Weather situation 1 attracted more booking which seems obvious as it is clear weather.
4. Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
5. When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
6. Booking seemed to be almost equal either on working day or non-working day.
7. 2019 attracted a greater number of booking from the previous year, which shows good progress in terms of business.

---

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

Solution: drop_first = True implies whether to get n-1 dummies out of n categorical levels by removing the first level.
It is important to use, as it helps in reducing extra variable created during dummy variable creation. This helps reduce the correlation created (multicollinearity) among dummy variables.

It is one of the parameters in pd.get_dummies function.
Parameter Type – bool
Default Value – False
Values Accepted – False, True

For Example – In our dataset we had season column (which had 4 values) which on applying get_dummies with drop_first = True created 3 dummy variables – season2, season3, season4. It would have created 4 dummy variables in case drop_first would have not been specified as True.

---

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

Solution: temp variable seems to have highest correlation with the target variable **cnt**. Graph seems to show quite good linear relationship.

---

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

Solution: Assumptions of Linear Regression after model building were done by validating below assumptions -

1. **Linear Relation Validation**
    a. Linear Relationship should be visible between dependent and independent variables.
2. **Independence of residuals**
    a. No Autocorrelation in the residuals.
3. **Homoscedasticity**
    a. No Heteroskedasticity. There should be no visible pattern in the residual values.
4. **Checking Multicollinearity**
    a. There should be insignificant multicollinearity between the variables.
5. **Normal Distribution of Error Terms**
    a. Residuals must be normally distributed.

---

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

Solution: 3 Top features which contributed significantly towards demand of shared bikes are :

1. Temp (coefficient value = 0.5804)
2. Weather Situation 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) (coefficient value = -0.2768)
3. Season 4 (winter season) (coefficient value = 0.1291)

General Subjective Questions :

1. Explain the linear regression algorithm in detail.

Solution : It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.
It analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means 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).

Mathematical Representation:

$y= a_0+a_1x+ \varepsilon$

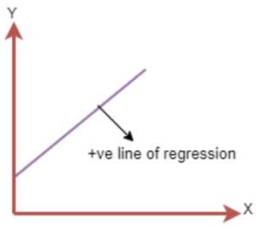
Here,
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

Types of Linear Regression :

1. Simple Linear Regression - single independent variable is used to predict the value of a numerical dependent variable.
2. Multiple Linear Regression – If more than one independent variable is used to predict the value of a numerical dependent variable.

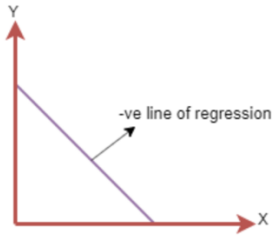
Regression line - Linear line showing the relationship between the dependent and independent variables. It can be of two types :

1. **Positive Linear Relationship:**
    a. If target variable on Y-Axis increases with increase in value of independent variable on X-Axis, then such relationship is called as Positive Linear Relationship.



The line equation will be: $\mathbf{Y= a_0+a_1x}$

2. **Negative Linear Relationship:**
    a. If target variable on Y-Axis decreases with increase in value of independent variable on X-Axis, then such relationship is called as Negative Linear Relationship.



The line of equation will be: $Y= -a_0+a_1x$

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values.

Following are the assumptions that are made on dataset by the linear regression algorithm :

1. **Multi-Collinearity** – It assumes very little or no significant multi collinearity in the given data. It generally occurs when the independent variables or features have dependency in them.
2. **Auto-correlation** - Assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
3. **Linear Relationship between variables** – Desires that the relationship between the dependent and independent variables is linear.
4. **Normal Distribution of Errors** -Desired error terms to be normally distributed.
5. **Homoscedasticity** – No visible pattern in residual values.

---

2. **Explain the Anscombe's quartet in detail.**

Solution: 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 when they are represented graphically. Each graph will look different even though statistics seem to say they are same.

This is how the dataset looks like:

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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

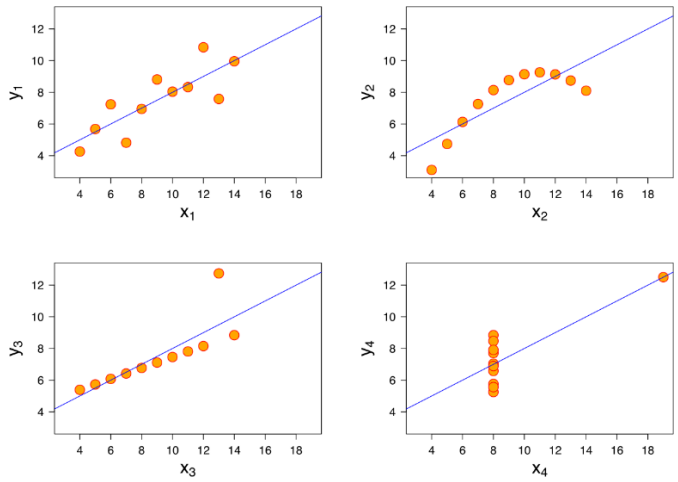
The Descriptive stats of above dataset looks like this:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

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 appear 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.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

---

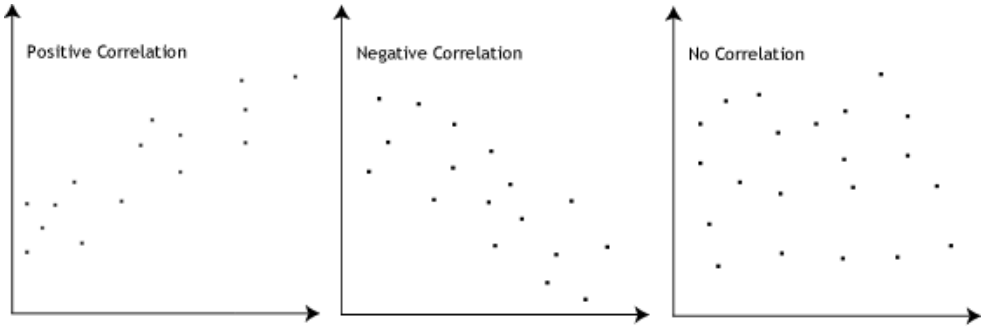
3. **What is Pearson's R?**

Solution : Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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.

It can take a range of values from +1 to -1
Values greater than 0 indicates positive association. That is increase or decrease in one causes increase or decrease in other respectively.
Values less than 0 indicates negative association. That is increase or decrease in one causes decrease or increase in other respectively.
Value equal to 0 says there is no association between the variables and are independent of each other.

Below graph illustrates the scenarios:



---

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Solution : Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It 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.
Example: If an algorithm is not using feature scaling method, then it can consider the value 5000 meter to be greater than 7 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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?**

Solution: VIF = infinity can only happen in the case of perfect correlation. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 6, this means that the variance of the model coefficient is inflated by a factor of 6 due to the presence of multicollinearity.
If **VIF is infinity that automatically means that there is a perfect correlation between 2 independent variables**. In this case, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing multicollinearity. Infinite VIF indicates that the corresponding variables can be expressed by a linear combination of other variables.

---

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

Solution: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample test.