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)**

From the analysis on categorical columns using the boxplot and bar plot, following points can be inferred from the visualization –

- Maximum bike booking were seen in fall **season**. This was followed by summer & winter. This indicates, season can be a good predictor variable.
- More bike bookings were happening in the **months** June, July, August, September and October with a median of over 4000 booking per month. Trend is seen to be increasing from beginning of the year till mid of the year (as maximum can be seen in mid months above) and then decreasing trend is observed towards year end. This indicates, month has some trend for bookings and can be a good predictor variable.
- Majority of the bike bookings were seen during clear **weather**, which was followed by misty weather. This indicates, weather situation does show some trend towards the bike bookings can be a good predictor variable.
- Bike bookings were more when it is not a **holiday** which indicates that holiday might not be a good predictor variable.
- **Weekday** variable shows very close trend, hence, this variable may have some or no influence.
- **Working day** shows close trend, hence, this variable may have some or no influence.
- It is observed that there is significant increase in the count of bike bookings in the **year 2019** when compared to 2018 with respect to all the categorical variables. So year can be a good predictor variable.

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

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and not B, then It is obvious that it is C. So we do not need 3rd variable to identify the C.
Basically- **drop_first=True**, creates k-1 dummy variables for k categories to avoid dummy variable trap in some of the machine learning models such as regression.

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

Pair-Plot shows strong linear relationship between 'temp' and 'atemp'. Hence, both parameters can't be used for model building due to multicollinearity. These variables 'temp' and 'atemp' 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)**

   Following assumptions of Linear Regression Model are verified:
   - **Linear relationship** between target variable cnt and dependent variable temp is seen in the pairplot.
   - **No or little multicollinearity** is verified as all the predictor variables have VIF value much less than 5.
   - **Normal Distribution of residuals**- for this histograms of the error terms are plotted and it is observed that its normal distribution.
   - Residuals' following a normal distribution is cross-verified by QQ plot.
   - **Homoscedasticity** is preserved as visible pattern in residual values is not observed.
   - **Independence of residuals**: To verify that the observations are not auto-correlated, Durbin-Watson test is conducted and the Durbin-Watson value for Final Model lr5 is 2.0877, which signify that there is almost no autocorrelation between observations.
   - **All independent variable are uncorrelated with error term** as no correlation is observed between any of the independent variables and the error term

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

   As per the final Model, the top predictor variables that influences the bike booking are:
   - **Temperature (temp)** - with coefficient value of '0.5709'
   - **Weather Situation (weathersit_light_snow)** - with coefficient value of '-0.2439'
   - **Year (yr)** - with coefficient value of '0.2294'

General Subjective Questions

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

   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**.
   Linear regression may be defined as the statistical model that 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).

   Mathematically the relationship can be represented with the help of following equation –

   $$Y = mX + c$$

   Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y, c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

We update m and c values to get the best fit line by minimizing **Cost Function:**

$$minimize\ \frac{1}{n}\sum_{i=1}^{n} Ypred - Ytrue)^2$$

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. To update m and c values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line, the model uses **Gradient Descent**. The idea is to start with random m and c values and then iteratively updating the values, reaching minimum cost.
Once we find the best m and c 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.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship: A linear relationship will be called positive if with increase in dependent variable, independent variable also increases.
- Negative Linear relationship: A linear relationship will be called negative if with increase in dependent variable, variable decreases.

Linear regression is of the following two types −

- Simple Linear Regression- **X:** input training data is univariate – one input variable (parameter)
- Multiple Linear Regression- **X:** input training data is multivariate – multiple input variable (parameter)

Assumptions of Linear Regression model−

- Linear regression model assumes Linear Relationship between target variable and predictor variables
- Normality of error terms – Error terms should be normally distributed
- Multi-collinearity –Linear regression model assumes that there is very little or no multi-collinearity among the predictor variables.
- Homoscedasticity –here should be no visible pattern in residual values.
- Auto-correlation – Linear regression model assumes is that there is very little or no auto-correlation in the residuals.

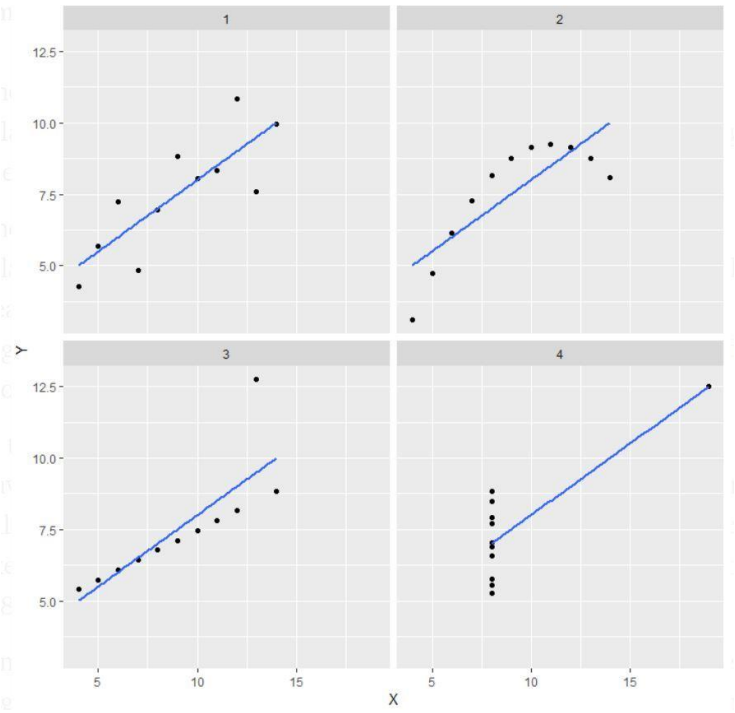
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
Below 4 sets of 11 data-points 11 data-points when analyzed using only descriptive statistics, the mean, standard deviation, and correlation between x and y were found same for all the 4 datasets.

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

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

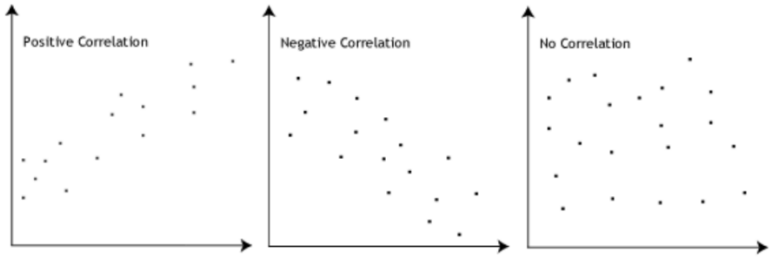
Explanation of this output:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point 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 analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

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 for multiple predictor variables. 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.

Feature scaling is about transforming the value of features in the similar range like others for machine learning algorithms to behave better resulting in optimal models. Feature scaling is required for multiple linear regression but not required for algorithms such as random forest or decision tree.

In multiple linear regression, rescaling is mandated for easy interpretation of coefficients. Different variables must be at comparable scale, so that coefficients obtained by fitting the regression model can be comparable during model evaluation.

Further, if the various variables are in the range of 0 and 1, then the optimization is much faster, as gradient descent algorithm running for minimization of cost function.

Scaling can be done using

- Normalization
- Standardization

Differences between these scaling techniques are as follows:

Si. No.	Normalization	Standardization
1	**Normalization** is about **transforming** the feature values to fall within the **bounded intervals (min and max)**	**Standardization** is about **transforming** the feature values to fall around **mean as 0 with standard deviation as 1**
2.	Normalizing data based on min-max scaling concepts. Minimum and maximum values of features are used for scaling.	The standardization technique is used to center the feature columns at mean 0 with a standard deviation of 1.
3.	More sensitive to outliers.	Less sensitive to outliers.
4.	Rescales the data set such that all feature values are in the range [0, 1]	Feature columns have the same parameters as a standard normal distribution
5.	Normalization is useful when the data is needed in the bounded intervals.	StandardScaler cannot guarantee balanced feature scales in the presence of outliers.
6.	$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$	$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$

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

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.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (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 this perfect multicollinearity.

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

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 tests.