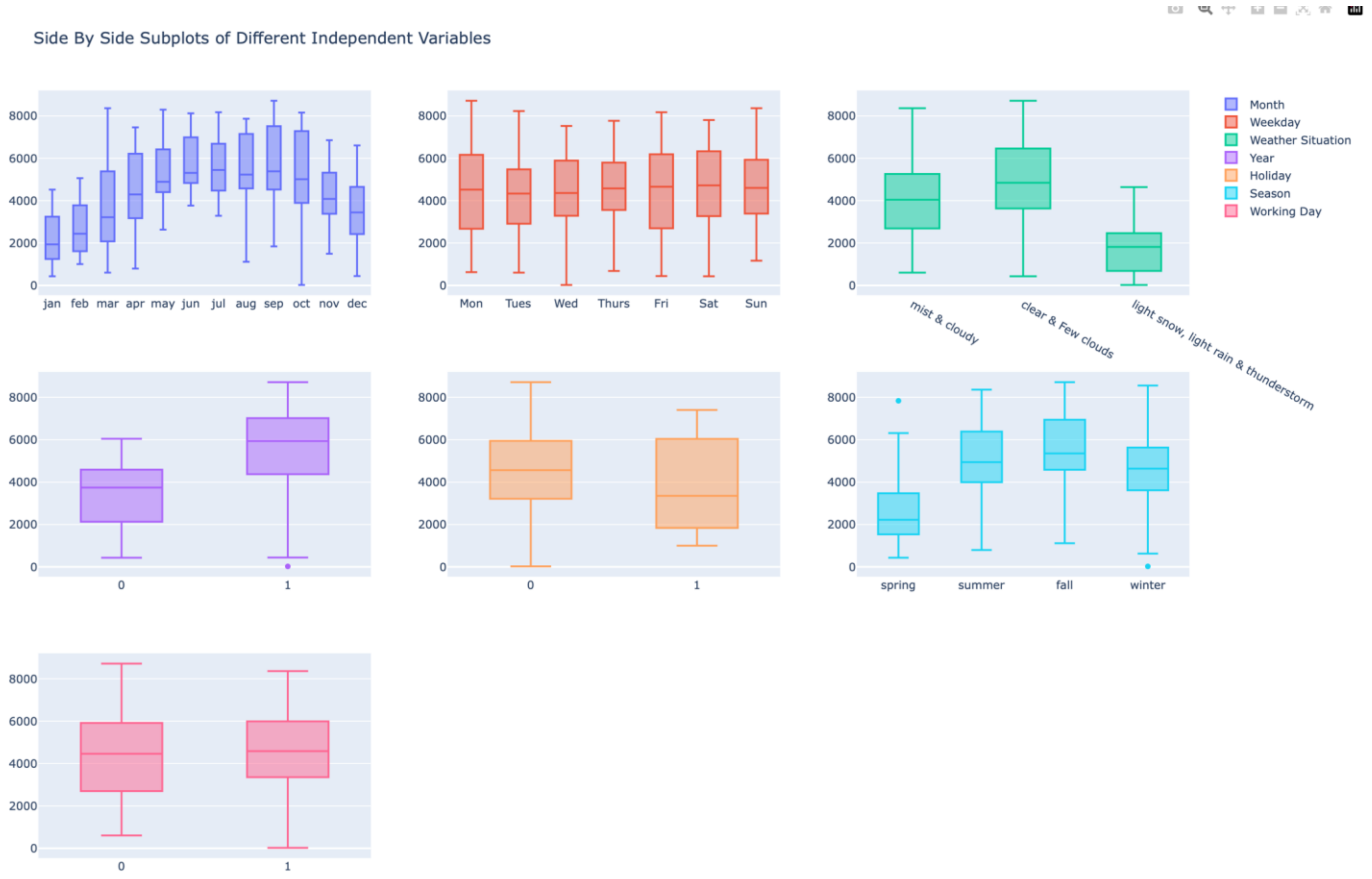
Assignment-based Subjective Questions

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



From the above chart following bivariate analysis can be drawn,

a) Most number of bikes are booked in the months from April to October; However top 2 months are September & October
b) Most number of bikes are booked on Monday, Friday & Saturday, and the median of all the days is close to 5000
c) Most number of bikes are booked during clear & few clouds season. With 75 percentile booking ~ 6500
d) In year 2009 more bikes are booked as compared to the year 2008; This means booking of the bikes are on the increasing trend/year
e) More bikes are booked on Working days. May be this attribute won't contribute towards the model analysis
f) Top 2 seasons for booking bikes are - Summer & Fall
g) During the spring season the number of booking falls
h) season, months, days, year, holiday all are contributing towards the booking of the bikes

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

It is used to drop the actual column from which the dummy columns are created to reduce the number of columns in the dataset, and this will also avoid the multicollinearity with the dummy n-1 columns created

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

a) Between cnt & temp (temperature in Celsius)
b) Between cnt & atemp (feeling temperature in Celsius)

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

a) Residual Analysis of the training data, which should be normally distributed
b) By verifying that the variance of the error terms is Homoscedasticity i.e. the spread of residuals should be constant for all values of X
c) Create a scatter plot between the fitted and actual values of 'count' using training data, which should be linear

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

a) Temperature – A unit increase in Temperature predictor increases the Bike booking numbers by 0.4509 unit

b) Year – A unit increase in Year predictor increases the Bike booking numbers by 0.2344 unit
c) Weather Situation - light snow, light rain & thunderstorm - A unit increase in this weather situation predictor decreases the Bike booking numbers by -0.2868 unit

General Subjective Questions

1. Explain the linear regression algorithm in detail

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 linear regression model the output variable is continuous in nature & it is a statistical method used in machine learning for predictive analysis. There are two types of linear regression simple and multiple linear regression.

A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line. The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

There are different ways to predict linear regression such as linear equation with R square statistics, MSE, Adjusted R square statistics. Along with the cost function, a 'Gradient Descent' algorithm is used to minimize MSE and find the best-fit line for a given training dataset in fewer iterations, thereby improving the overall efficiency of the regression model.

For simple linear regression the linear equation formula is $y = \beta 0\ X + c$, where

- y = output variable. Variable Y represents the continuous value that the model tries to predict
- X = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable X represents the input information provided to the model at any given time
- c = y-axis intercept (or the bias term)

- m = the regression coefficient or model parameter. In classical statistics, m is the equivalent of the slope of the best-fit straight line of the linear regression model

Assumptions of simple linear regression:

- Linear relationship between X and y
- Normal distribution of error terms
- Independence of error terms
- Constant variance of error terms

Building a linear model

- OLS (Ordinary Least Squares) method in statsmodels to fit a line
- Summary statistics
  - F-statistic, R-squared, coefficients and their p-values

Residual Analysis

- Histogram of the error terms to check normality
- Plot of the error terms with X or y to check independence

Predictions

- Making predictions on the test set using the 'predict()' function

A multiple linear regression model represents the relationship between two or more independent input variables and a response variable. Multiple linear regression is needed when one variable might not be sufficient to create a good model and make accurate predictions.

For multiple linear regression the linear equation formula is y = β0 X + β1 X + c, where everything remains same as simple linear regression; however, we have more than one features or independent variables. It is used when a lot variance isn't explained by just one feature.

There are few more considerations to be made for multiple linear regression model

- Overfitting - When the model becomes complex and gives very good results in training data and fails in the testing data
- Multicollinearity - To identify if there is any dependency within the pool of independent variables to remove redundancy
- Feature selection - Out of the pool of many features what features are most important. We drop the redundant features and those features that are not helpful in prediction

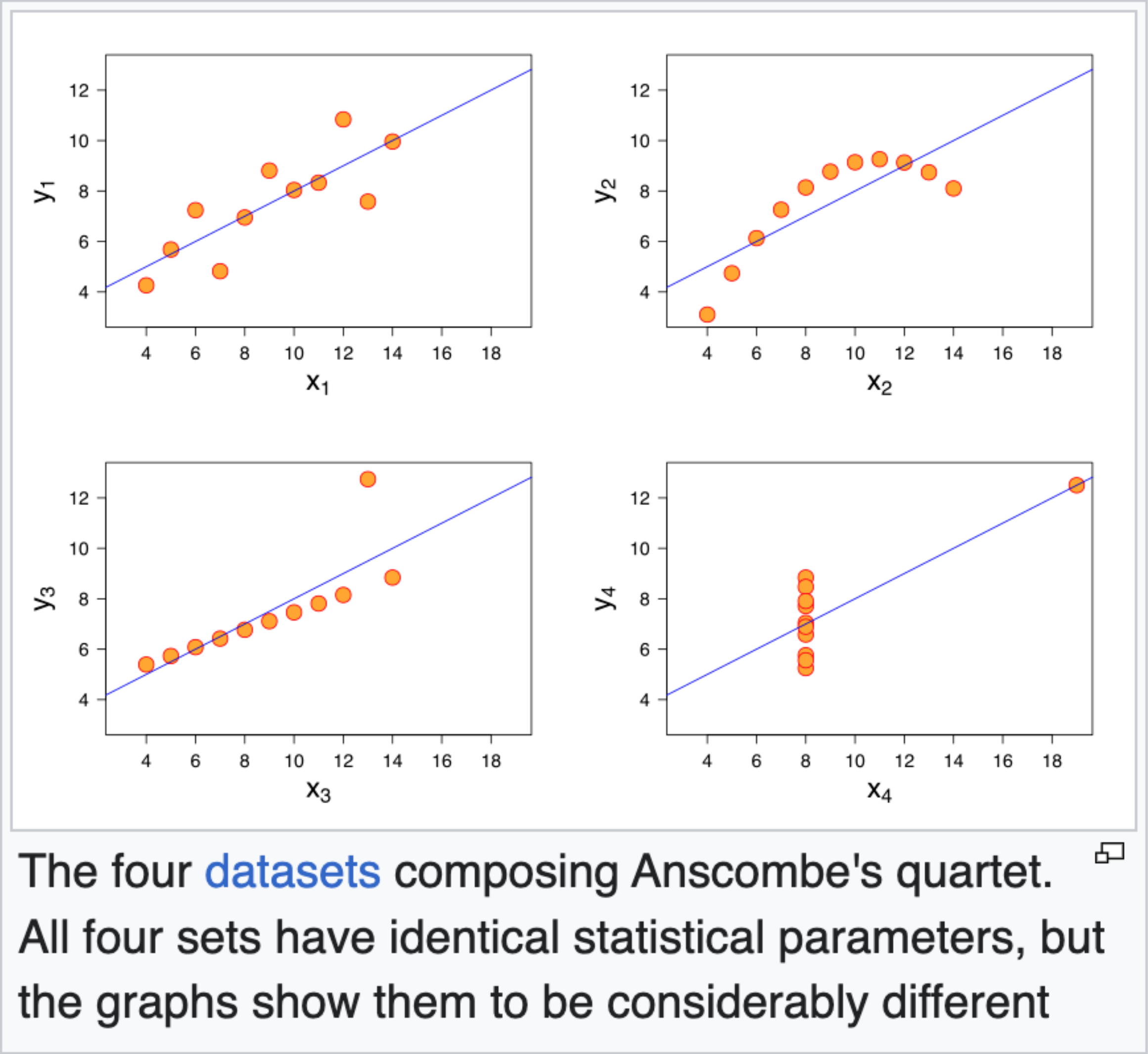
Assumptions & Residual Analysis definition remain same for multiple linear regression model

2. Explain the Anscombe's quartet in detail

Sometimes, even the statistical information summarised from the data may mislead to wrong conclusions. Therefore, we should visualise the data often to understand how different features are behaving.

Even if we calculate the standard deviation or mean of those different datasets, we may get the same results of both the variables, and this will not help us to conclude the analysis from the dataset as all the datasets are showing the same information.

To better understand the impact of input variable to output variable we should plot scatter graphs for all the datasets. This will help us to visualize the graphs and get appropriate analysis. An example is shown below,



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

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

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,

Benchmark for Pearson's R is described below,

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all the following are true:

- **Both variables are quantitative:** We will need to use a different method if either of the variables is qualitative
- **The variables are** normally distributed**:** We can create a histogram of each variable to verify whether the distributions are approximately normal
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers
- **The relationship is linear:** means that the relationship between the two variables can be described reasonably well by a straight line

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

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

Scaling is used for:

a) **Ease of Interpretation** specifically in case of multiple linear regression because features could be at different scale due to which model parameters/weights do not specify the importance of the that feature on the whole dataset. With scaling we can compare the one coefficient with the other as they are on the same scale
b) **Faster convergence** for gradient descent methods, as the values across all the features are close to each other or on same scale

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.

There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all the data in the range of 0 and 1, if there are lot of outliers in the data then Standardisation scaling should be used as MinMax scaling will scale outliers in the range of 0 and 1

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

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

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

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 means VIF is very high and perfect correlation exists among independent variables.

We need to drop one or more of the predictors from the dataset which is causing this perfect multicollinearity which will further reduce the VIF

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

A Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution or uniform distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line. In other words, whether the two datasets have similar distributions.

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, we can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. We can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, we need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution or uniform distribution.

A Q-Q plot helps to compare the sample distribution of the variable at hand against any other possible distributions graphically.