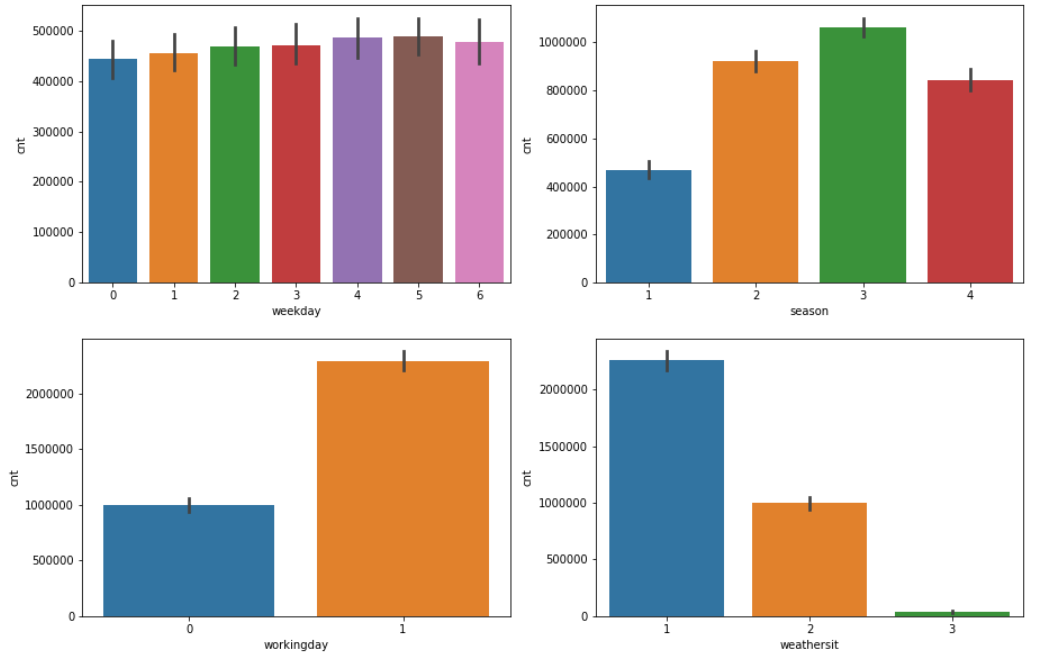
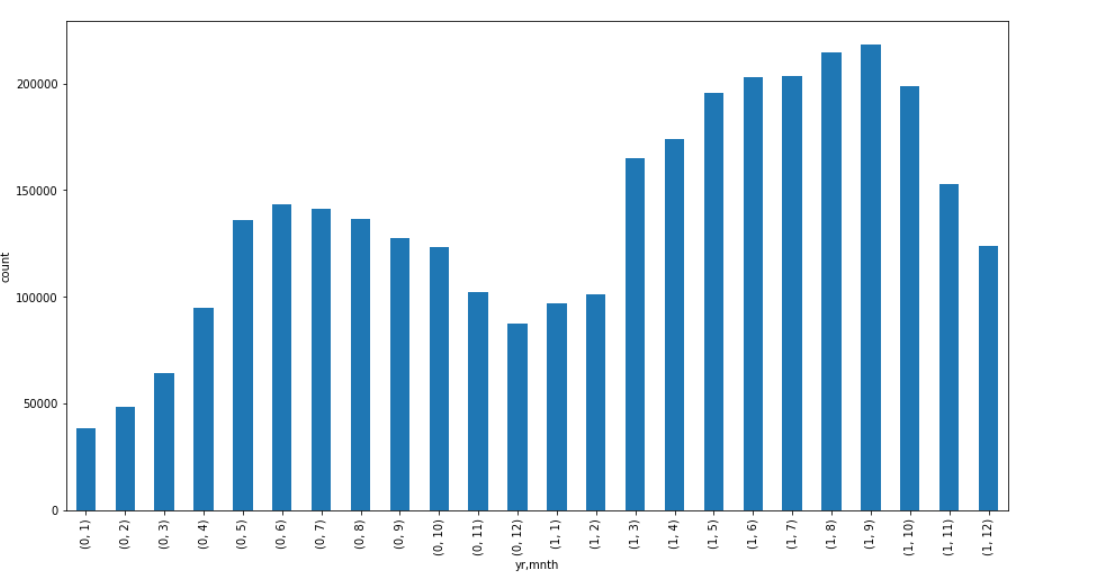
**Assignment Subjective Questions**

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

**A1.** Key Observations from categorical variables are as follows

* During Fall (season), highest number of bookings are recorded.
* Bookings are very high on working days then on non-working days. People prefer to use bike service for travelling to work.
* People are more likely to use bike service on a clear or partly cloudy day. and hardly use the service during snow season.
* From the last figure below, it is clear that over the two years the demand for bikes has increased
* This also confirms that during winters i.e. months January, February, November, December demand decreases significantly.



**Q2. Why is it important to use drop\_first=True during dummy variable creation?**

**A2.** If we don’t use drop\_first=True we will get a redundant variable. For example, in our data set.

Holiday column is a categorical column.

|  |  |
| --- | --- |
| **Row** | **Holiday** |
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |

After applying get dummies without drop\_first

*pd.get\_dummies(df.Holiday, prefix=’Holiday’)*

|  |  |  |
| --- | --- | --- |
| Row | Holiday\_0 | Holiday\_1 |
| 0 | 1 | 0 |
| 1 | 1 | 0 |
| 2 | 0 | 1 |

Both the variables are clearly strongly correlated leading to multi-collinearity.

In general, among k categories – kth category can always be shown by setting k-1 categories as 0 or absent.

In context of Linear Regression Model, multi-collinear parameters are not advisable as it affects the interpretation and inference derived from the model.

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

**A3.** Temperature (temp or atemp) is strongly correlated with target variable **cnt.**

(Both atemp and temp are highly correlated among themselves so both can be considered)

Structure of plot for temp and atemp is also similar w.r.t. target variable count.

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

**A4.** Assumptions of linear regression are as follows:

* Linear relationship between X and Y variables.
* No multicollinearity among predicting variables
* Error terms are normally distributed.
* Error terms are independent of each other.
* Error terms have constant variance. (homoscedasticity)

Verification of assumptions on data set.

1. **Linear relationship between X and Y**.- We have plotted pair plots for numerical variables. It is quite evident that numerical variables have linear relation with target variable with different levels of variances.
2. **No multicollinearity among predicting variables**- This can be detected in 2 ways:
   1. **Scatter Plots** – We can visually detect any relation between the variables.
   2. **Variance Inflation Factor (VIF) -** VIF = 1/(1-R2). It is calculated by building model among predicting variables and determine if any variable can be predicted by other variables.
      1. **VIF<5 -** Variable is sufficiently independent
      2. **VIF>5:** Variable may or may not be independent. Worth further analysis
      3. **VIF>10:** Variable is not independent of others and should be either treated or eliminated.
3. **Error Terms are normally distributed –** Weperformresidual analysis on predicted values in training dataset. Plotting a dist plot can help in visually verifying this.
4. **Error terms are independent of each other –** Thiscan be again done using following ways
   1. **Scatter plots -** Thereshouldbe no visible pattern in the error terms.
   2. **Durbin Watson Test-** Test statistic obtained in this test has value from 0-4
      1. If value = 2 - no autocorrelation
      2. 0 to <2 - positive correlation
      3. >2 – 4 - negative correlation
5. **Error terms have constant variance. (homoscedasticity) –** Thiscan be again identified from scatter plot of error terms. The variance in terms should not change across different subsamples. The Goldfeld-Quandt Test can also be used to test for heteroscedasticity.

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

**A5.** Top 3 features contributing to explain the demands of shared bikes are:

1. weathersit\_3
2. temp
3. season\_spring (season\_1)

Year variable has a higher coefficient than temp. But the data is only available only for 2 years. Also, due to covid-19 the current year is seeing decrease in demand. If business expands after this period, the positive coefficient of year may look justified. If the covid-19 extends the demand may remain low and year may not be appropriate predictor. Hence, I think though over last 2 years demand had increased, considering the year variable for future predictions is not a good idea. In the notebook, I have experimented building a model by removing the ‘yr’ variable.

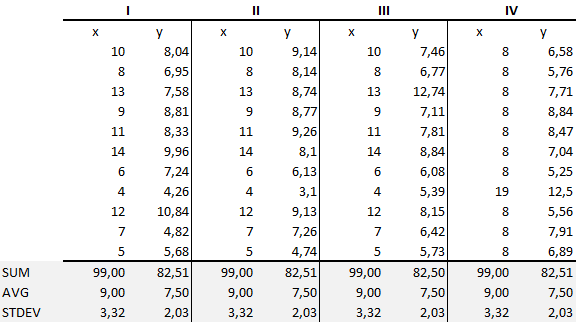
**General Subjective Questions**

**Q1. Explain the Linear Regression algorithm in detail.**

**A1.**

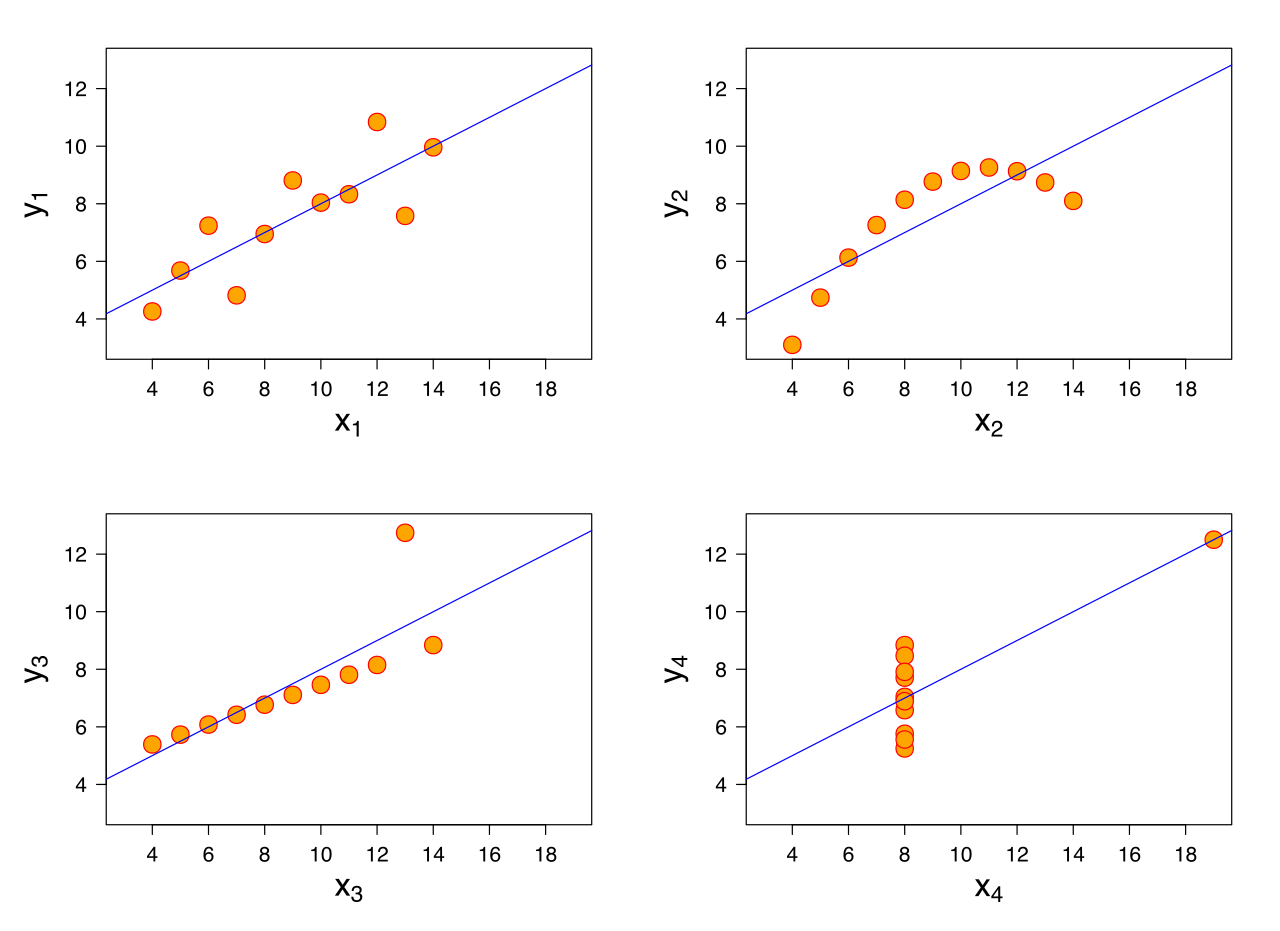
**Q2. Explain the Anscombe’s quartet in detail.**

**A2.** Anscombe’s quartet was developed by statistician Franscis Anscombe. It helped explain the concept that only mathematical value like mean, variance, standard deviation is not enough to get complete picture of the data. Visual analysis plays important role in understanding of the data.

It consists of 4 different data sets with exactly same statistical values. However, when these data sets are graphed, they show a totally different picture.

Source: [Dataset](https://miro.medium.com/max/1166/1*JyDU5qgFA-S2XOFBah9YcQ.png)

Source: [Anscombe Data Set Plot](https://miro.medium.com/max/1400/1*lZAlx4qLO4cHc5MERinBiw.png)



**Observations:**

1. Dataset 1 is clean and well fitting for linear models
2. Dataset 2 cannot be analyzed using linear regression due to curvature of the plot. It does not justify the first assumption of linear regression.
3. Dataset 3 – It has linear distribution largely. But it is the outlier which resulted in mathematical values matching the above 2 graphs
4. Dataset 4 – this is like dataset 3. If outlier is not present the values for statistical properties will be entirely different. Presence of outlier is sufficient for producing high correlation coefficient.