1. Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed.

In order to minimize the garbage that gets into our model, we have to preprocess the data. If we process directly to mine data that has not been preprocessed, we will face incomplete, and noisy databases. For example, the databases may contain:

* Fields that are obsolete or redundant,
* Missing values,
* Outliers,
* Data in a form not suitable for the data mining models,
* Values not consistent with policy or common sense.

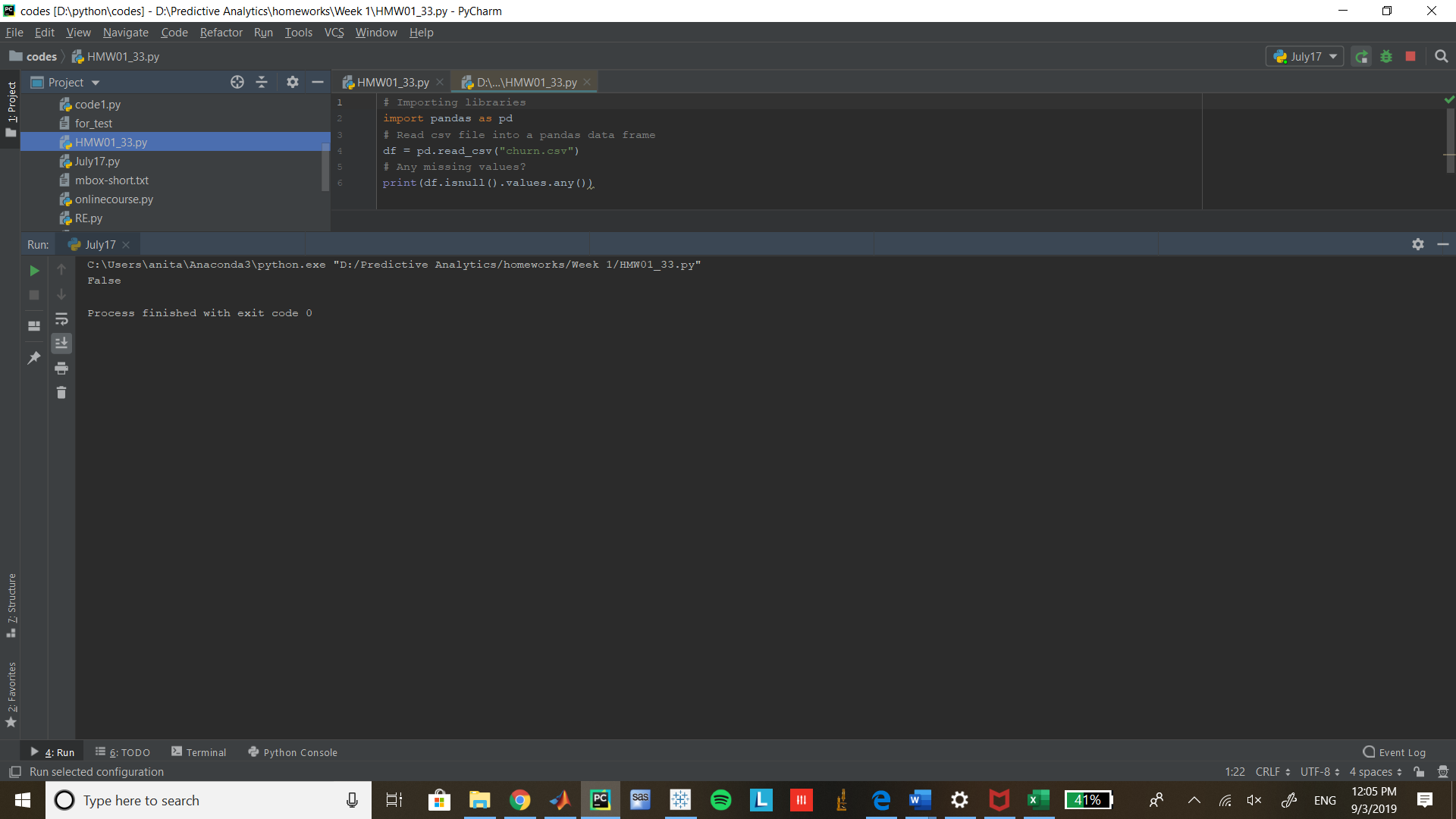
So, all outdated or incorrect information is still in dataset and our model productivity and quality will be decreased because of that.

3- Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.

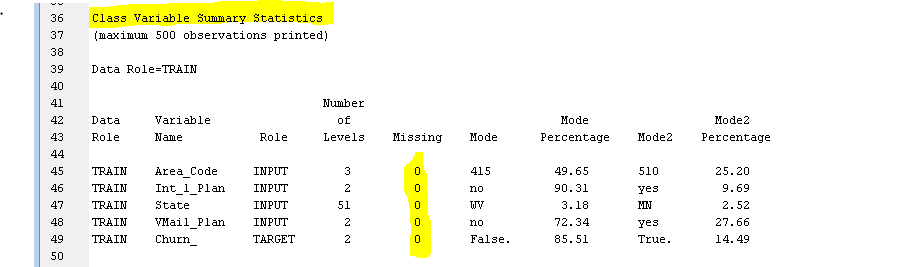
Omitting the records may be dangerous, since the pattern of missing values may in fact be systematic, and simply deleting the records with missing values would lead to a biased subset of the data. Further, it seems like a waste to omit the information in all the other fields, just because one field value is missing.

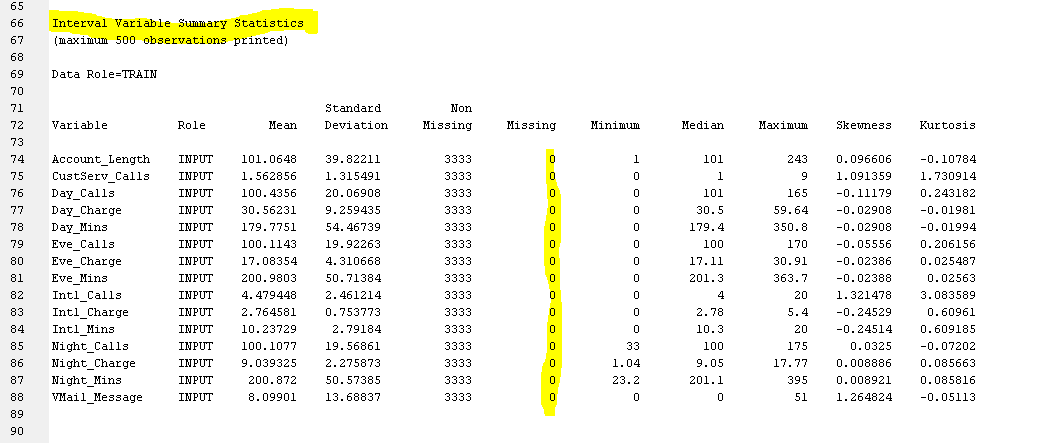
33-Explore whether there are missing values for any of the variables.

I used Python to see there is any missing value in dataset or not. The answer is False



Also, we can use SAS to check that thorough statexplore results. both class and interval variables have no missing value.

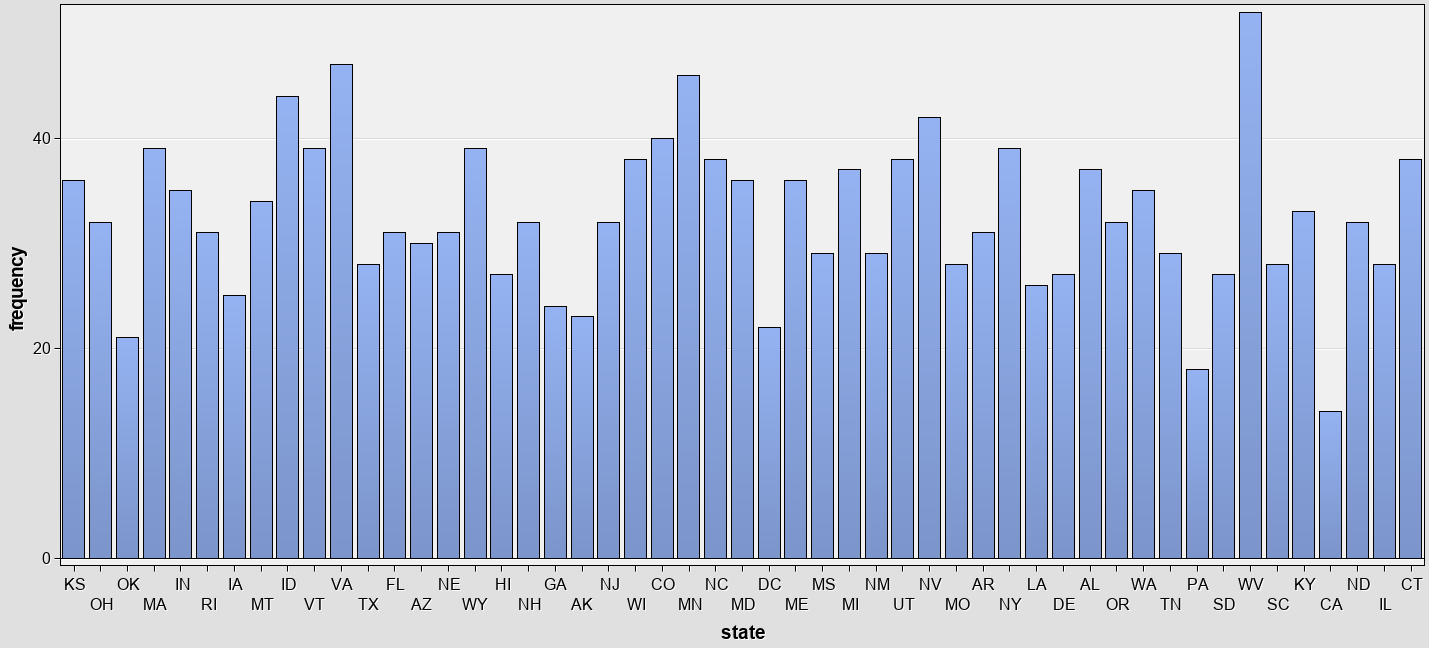




34- compare the area code and state fields. Discuss any apparent abnormalities.

for abnormalities in state variable: since state is a categorical variable, I used

bar chart to show States in graph.

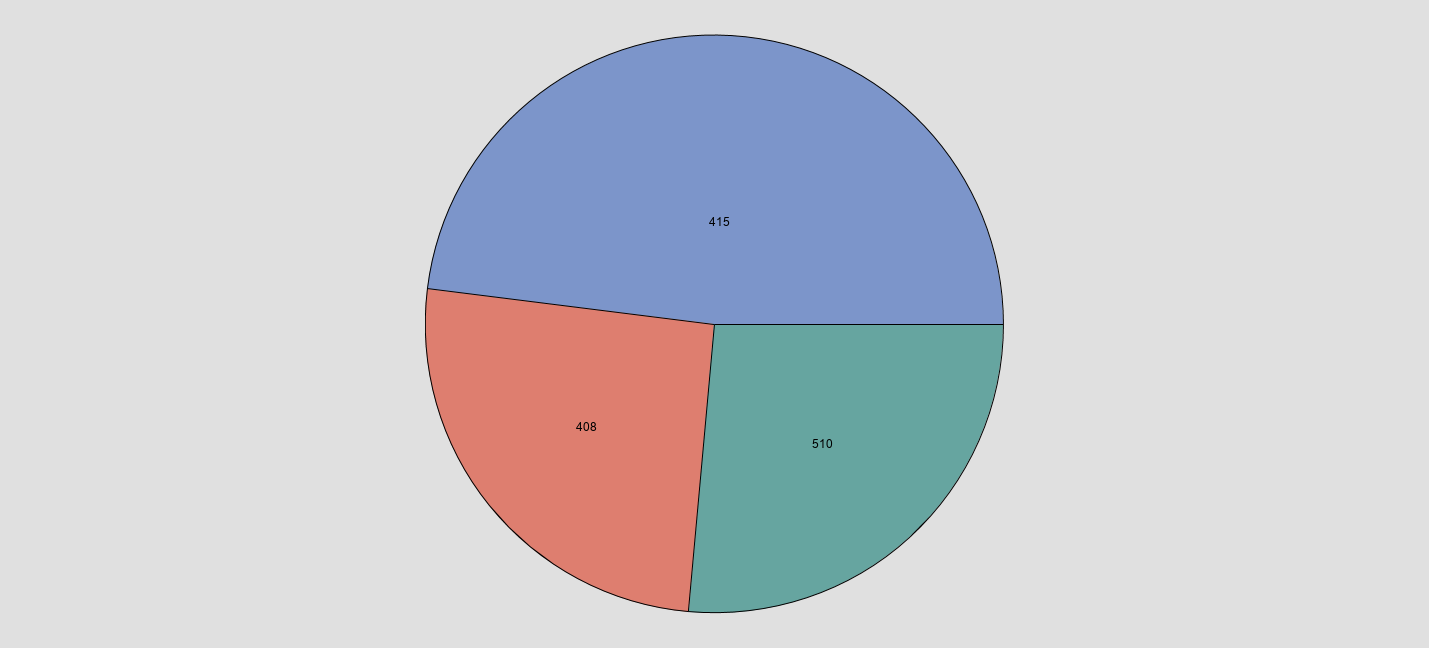


All the state names are two lettered and represent valid states of USA so, there

There are no abnormalities.

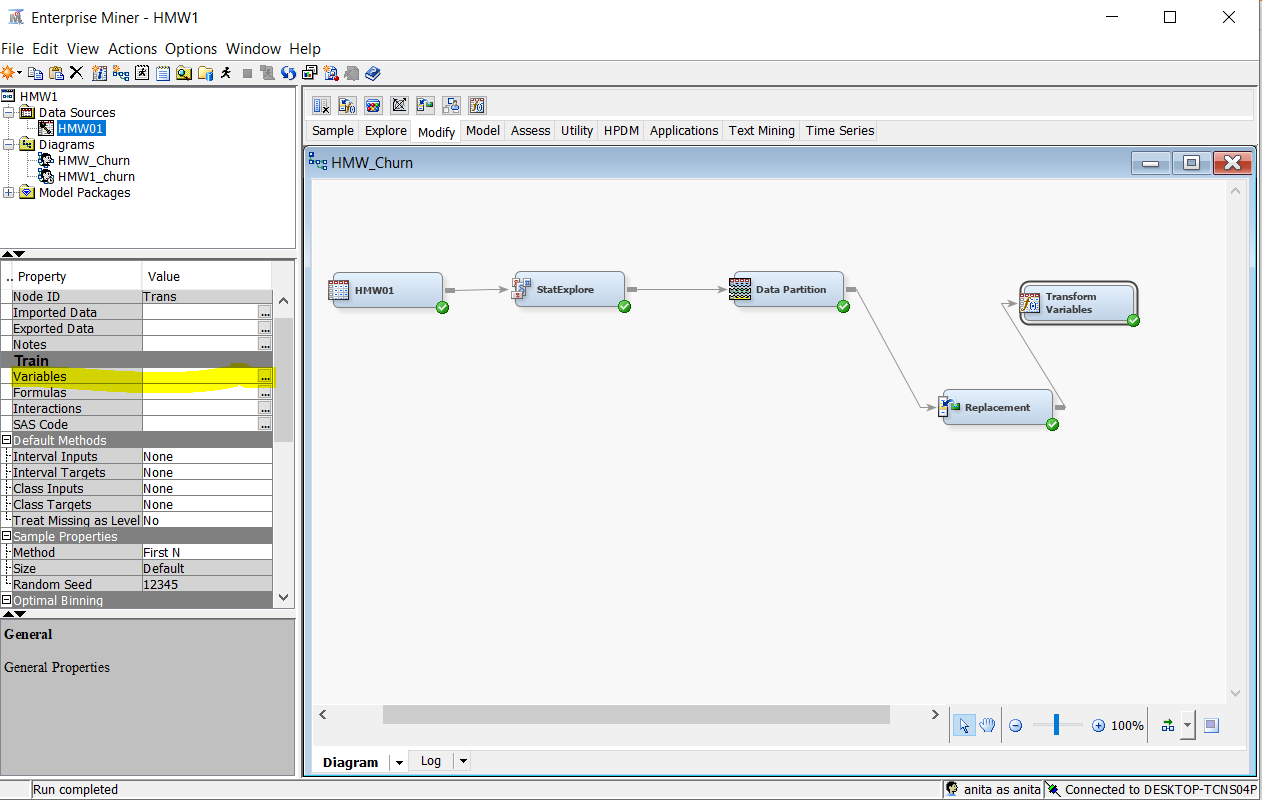
For abnormalities in area codes, we treat the values as categorical ones so I use

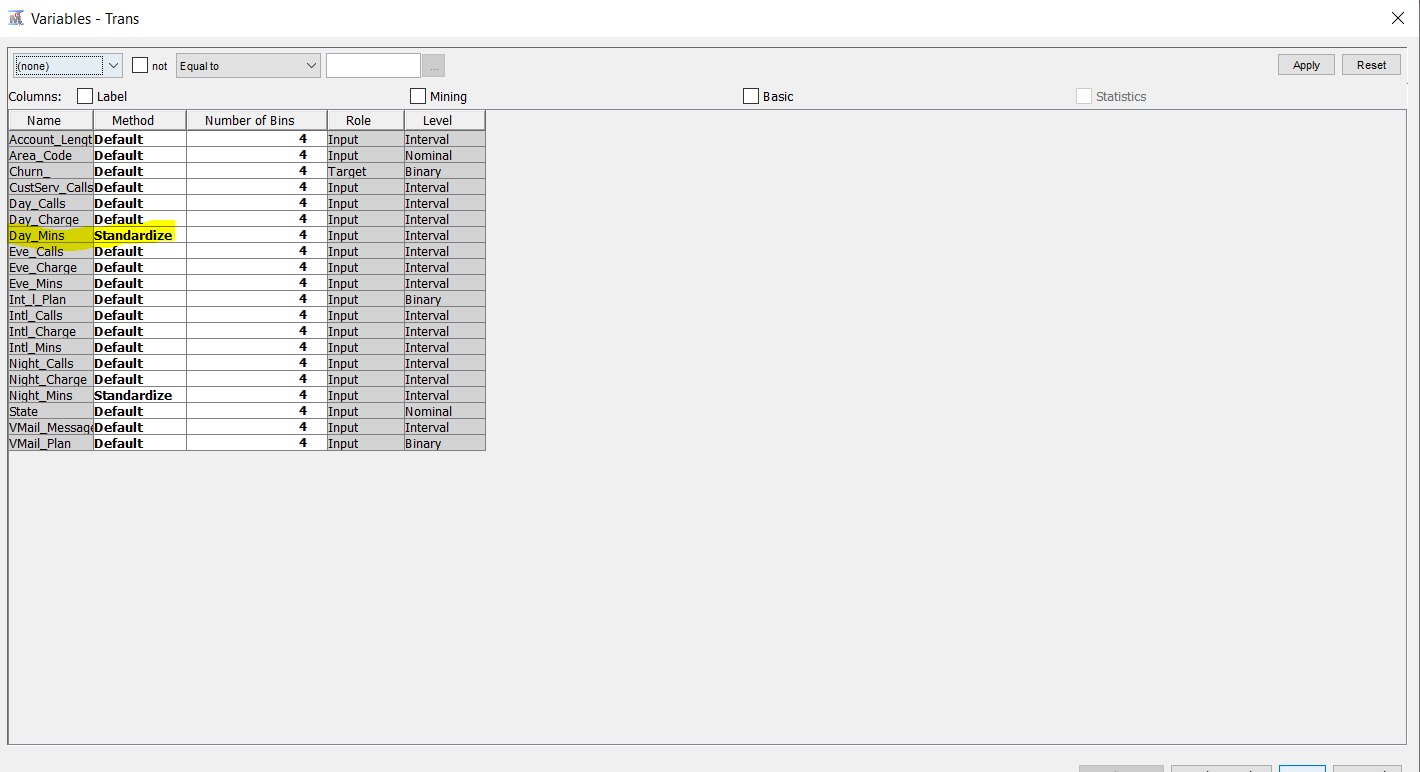
Pie-chart to show them on graph.

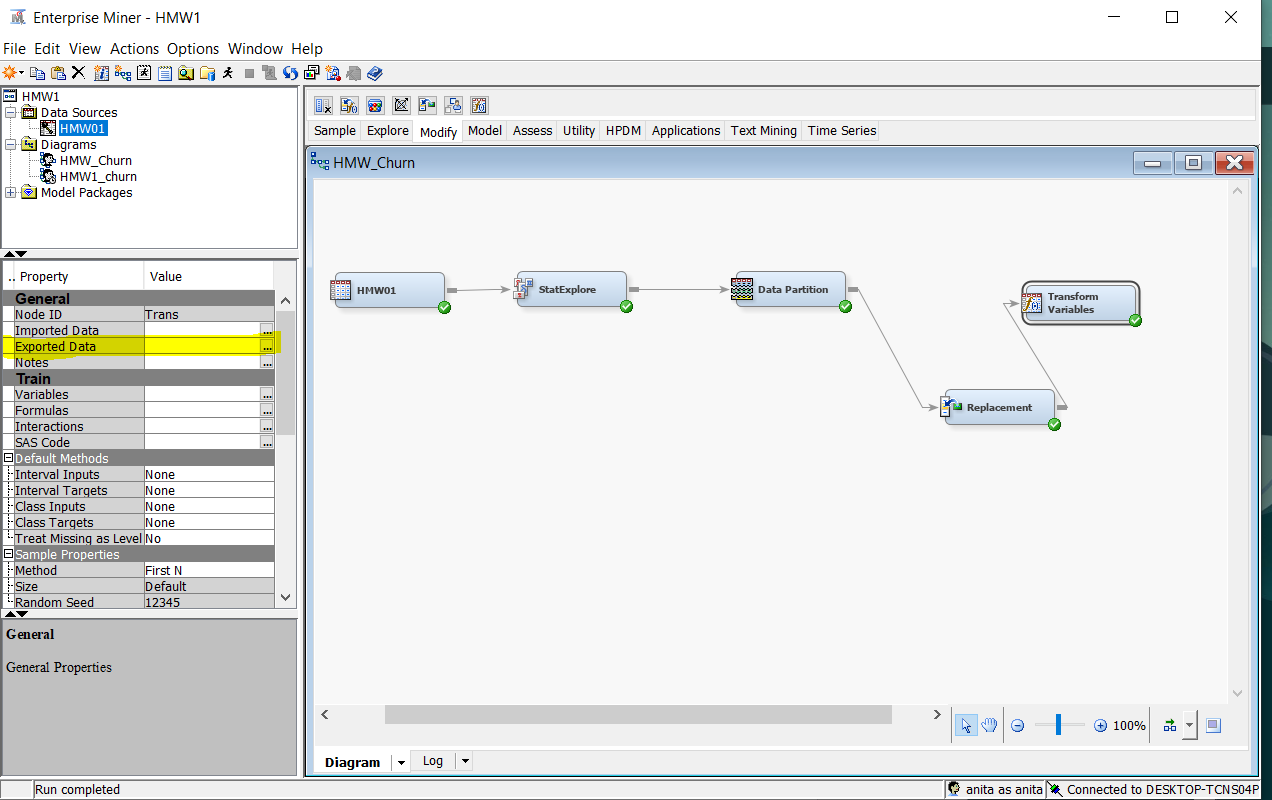


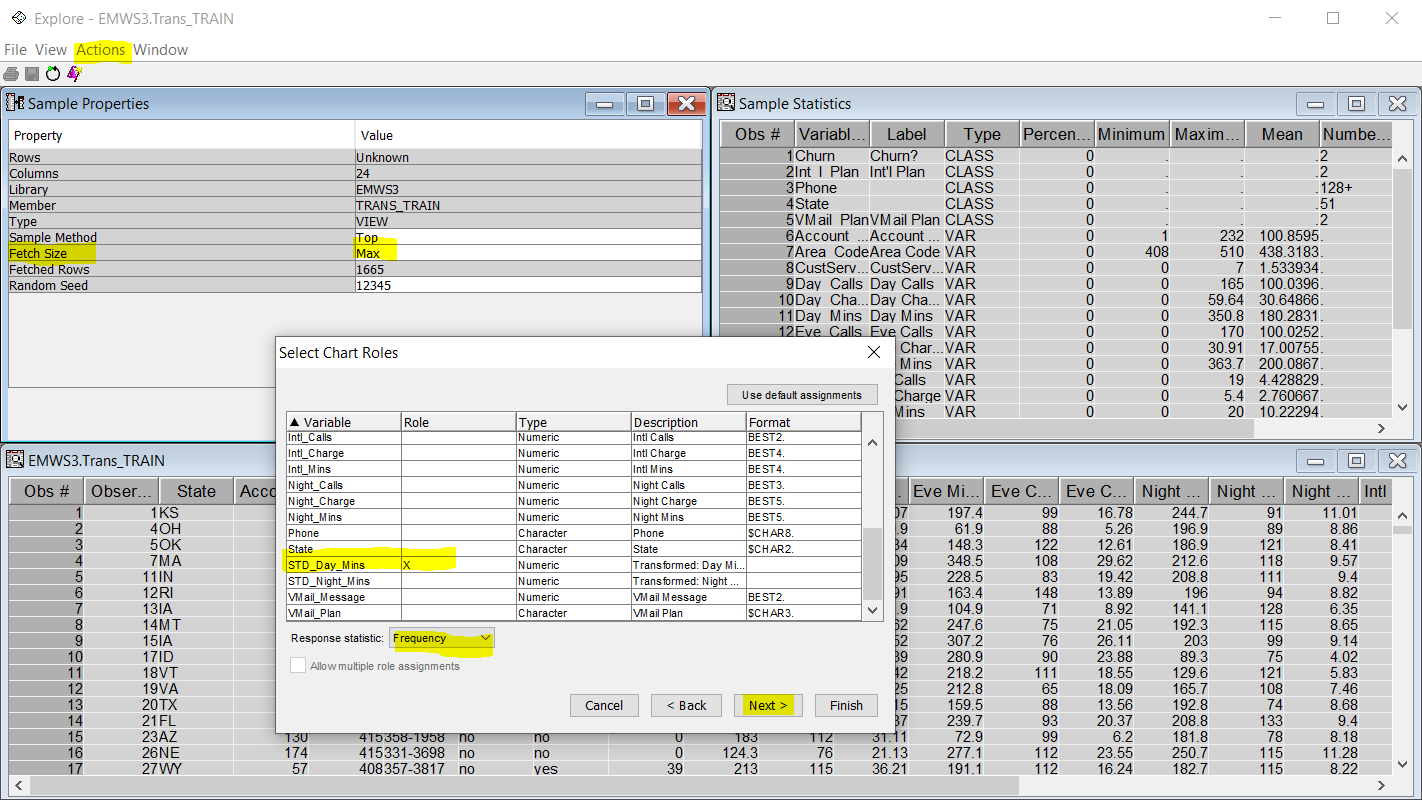
As It is drawn from chart, all states just assigned to 3 numbers 408(San Jose),415(San Francisco) and 510(Oakland) which is the abnormality here.

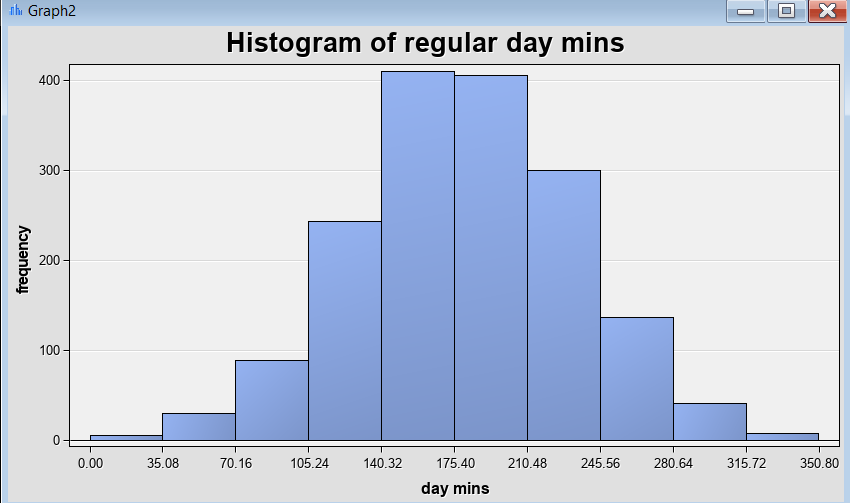
37 - Transform the day minutes attribute using Z-score standardization.

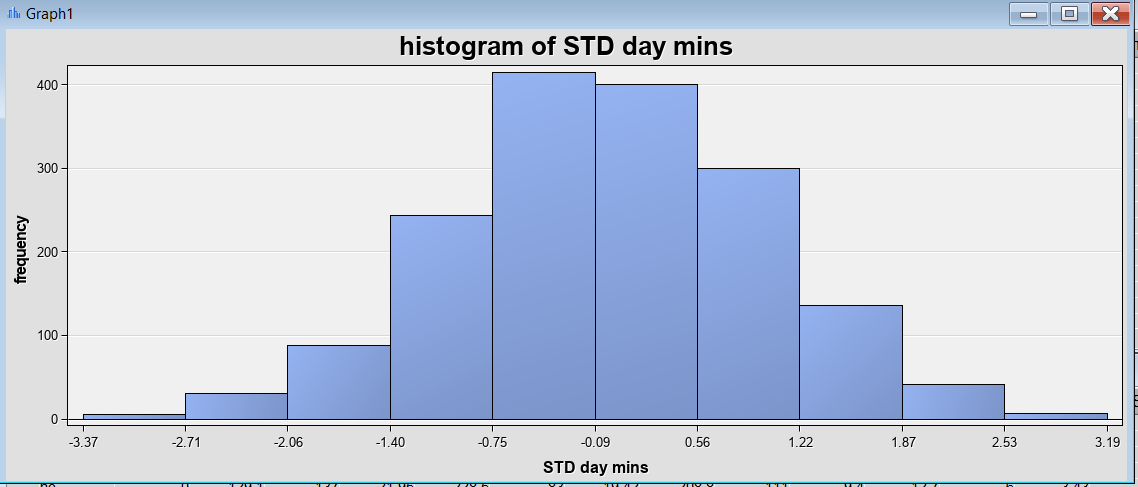






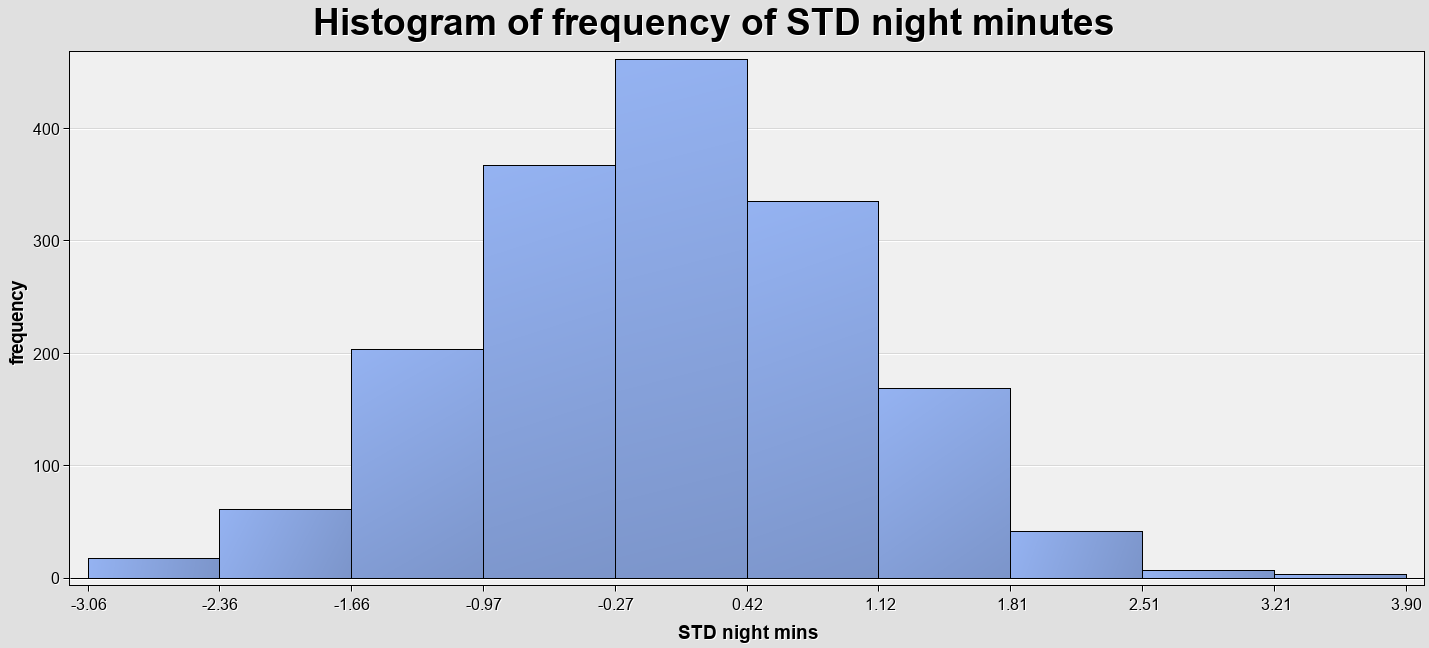


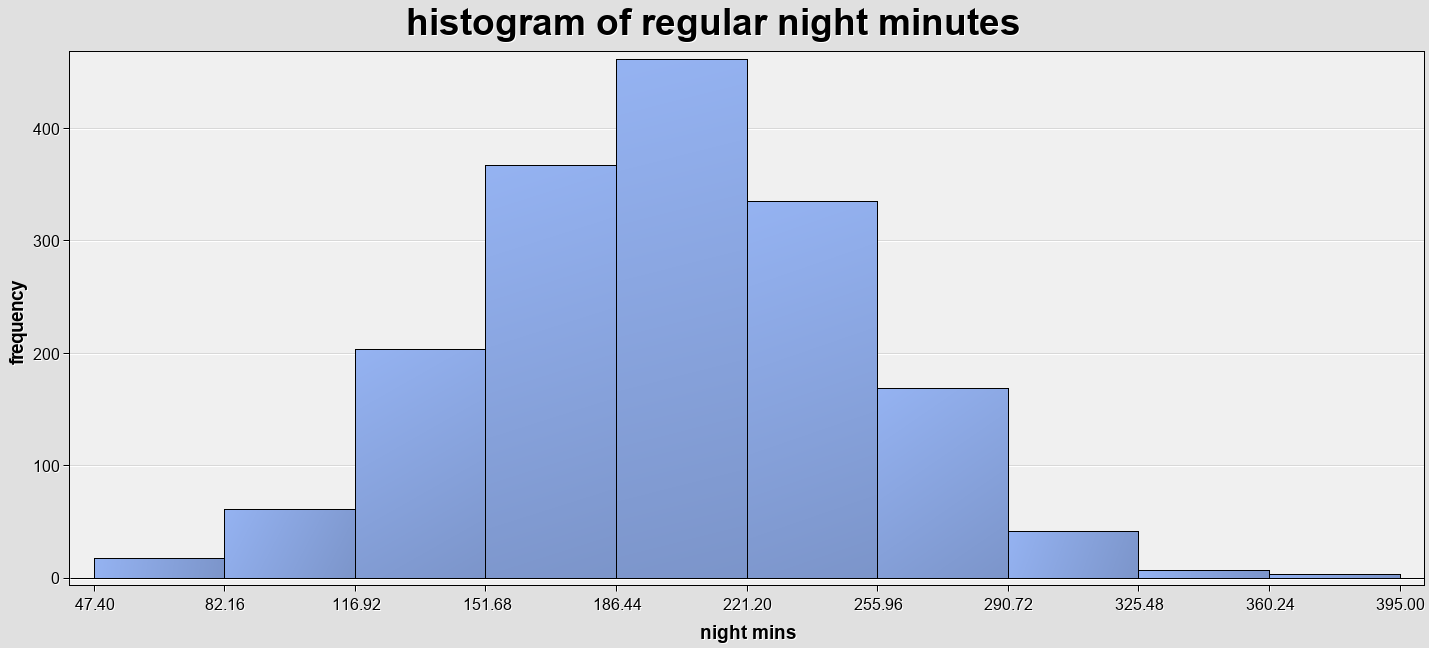




41 - Transform the *night minutes* attribute using *Z*-score standardization. Using a graph, describe the range of the standardized values.

Use the same method of drawing histogram as previous question





The range has been changed from [47.40 , 395.00] to [-3.06 , 3.90] which the mean value is changed to ˷ zero.