Course Name: Data Science: Process and Tools (DATASCI400)

1. **Project Problem**: Prepare a data set for machine learning.
   1. Model Type: none – just cleaning up the data
   2. Data Source: UCI Heart Disease
   3. URL : data (url: <http://archive.ics.uci.edu/ml/datasets/heart+Disease>)
   4. Final Conclusion/Result: outliers and aberrant data were removed. Values were normalized to 1. Categorical data was binned. New categorical variables were constructed. The old categorical data was removed
      1. Methods: use mean +/- 2 standard deviations to remove outliers
      2. Methods: remove NAs
      3. Normalize data (value – mean)/standard deviation
      4. Binned data turned to categorical variables
      5. Evaluation: goals achieved
2. **Project Problem**: Predict heart attacks using heart disease data
   1. **Model Type**: KNN and SVM
   2. Data Source: UCI Heart Disease
   3. URL : data (url: <http://archive.ics.uci.edu/ml/datasets/heart+Disease>)
   4. Final Conclusion/Result: ﻿SVM achieved better results than KNN and Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | SVM | KNN | Random Forest |
| Accuracy rate | 0.838 | 0.815 | 0.778 |
| Error rate | 0.161 | 0.161 | 0.161 |
| precision | 0.79 | 0.72 | 0.67 |
| Recall | 0.71 | 0.7 | 0.7 |
| F1 score | 0.75 | 0.74 | 0.68 |
| AUC | 0.9 | 0.88 | 0.81 |

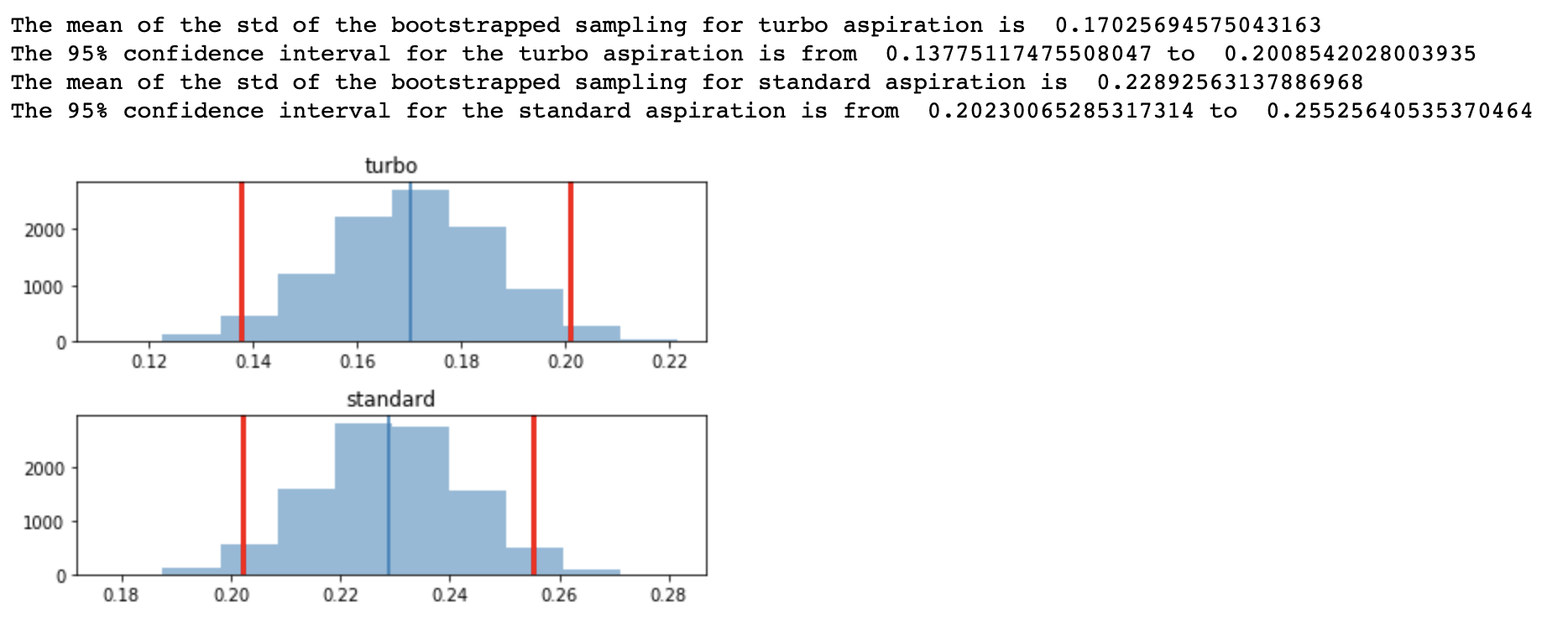
Course Name: Data Science 410: Methods for Data Analysis

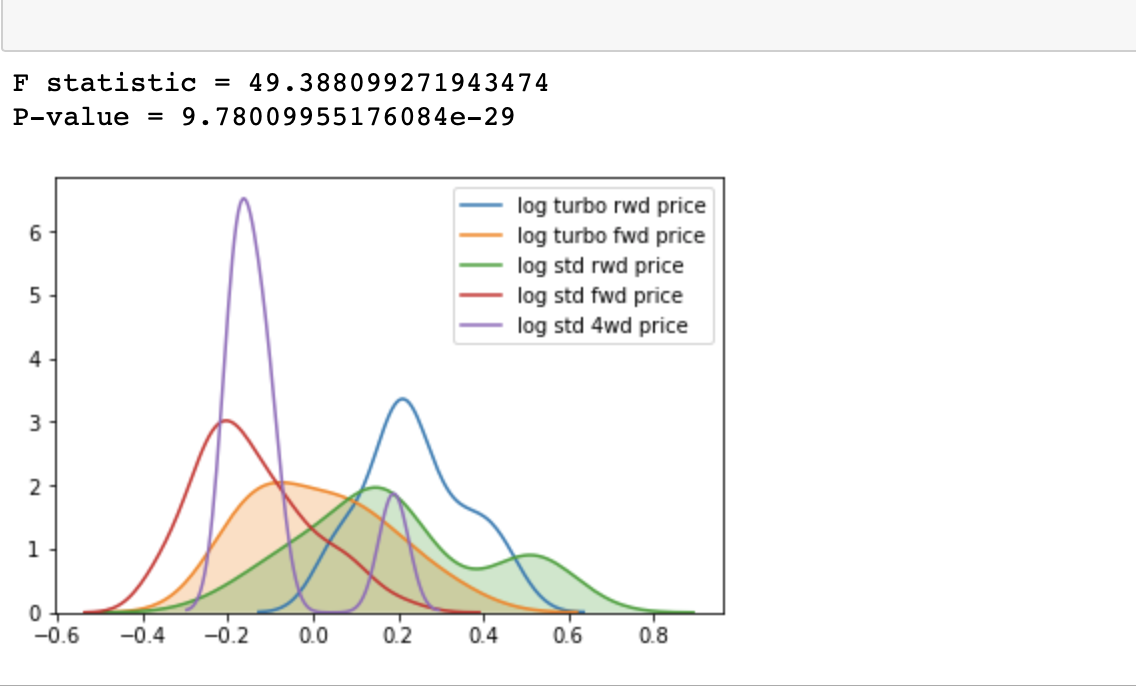
1. Data Visualization Complementary Views

1. Model Type: Visual exploration of the data for trends
2. Data Source: Jitter Head Count data set
3. url to the data set: <https://library.startlearninglabs.uw.edu/DATASCI410/Datasets/JitteredHeadCount.csv>
4. Final Conclusions: It appears that Craps, Big Six, and Texas Hold 'em are the most popular games. The table management (opening and closing) looks to be well-aligned with demand in Craps and most of the other games. But with Big Six and Texas Hold 'Em there are often many more tables opened than there is demand. These games need to be managed better. They appear to be especially in need of better management around noon.

2. Hypothesis Simulation:

1. Model type: hypothesis testing using
   1. Classical tests
   2. Tukey’s HSD
   3. Bootstrap Methods
   4. Bayesian model
2. Data source: UCI Machine Learning Repository – Automobile Data Set
3. url: <https://archive.ics.uci.edu/ml/datasets/automobile>
4. Final Conclusions: Logs of pricing show turbo and standard aspiration differ

Using Tukey’s Honest Significance we see



The log of the turbo versus standard aspiration pricing is significant and we can reject the null hypothesis that the pricing is unrelated to aspiration.

1. Regression Simulation:
   1. Model type: time series analysis
      1. Is the time series stationary
      2. Is there a significant seasonal component
      3. Forecast production for 12 months
   2. Data source: California Dairy Farmer Dataset
   3. url: <https://library.startlearninglabs.uw.edu/DATASCI410/Datasets/CADairyProduction.csv>
   4. Final conclusions/results

The series was not stationary: the autocorrelation died out slowly and the partial autocorrelation is significant for more than the first term. There was a pronounce seasonal component. Using Dickey-Fuller, the p-value was small that we can reject the null hypothesis that the residuals are not stationary. There was both autoregressive and Moving Averages behavior. The best model fit was using ARMA(2,1) with statistics of

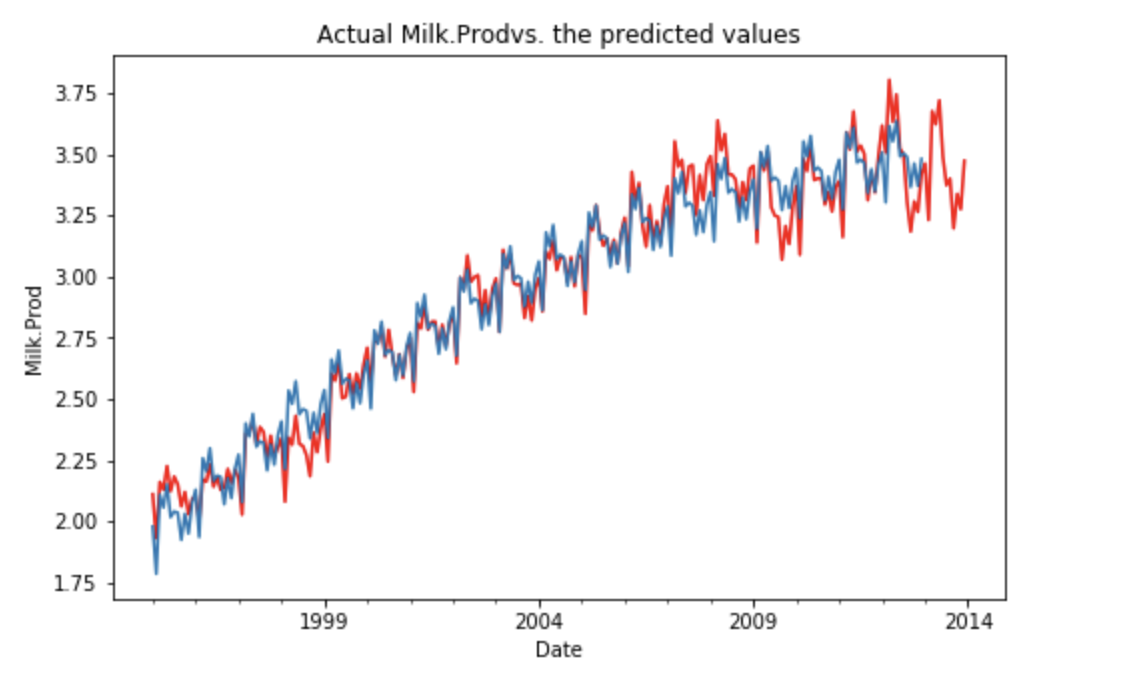
Forecast: [-0.00368767]

Standard Error: [0.01127966]

95%% Confidence Interval: (-0.0257, 0.0184)

The RMSE error was 0.277

A comparison of the model predictions versus the actuals (starting in 2013)



1. Independent Project: King County House Sales
   1. Model type: linear model
   2. King County House Sales
   3. url source: <https://library.startlearninglabs.uw.edu/DATASCI410/Datasets/kc_house_data.csv>
   4. Final results/conclusions: A linear model using the variables ‘month’, ‘day’, ‘waterfront’ ‘view’ grade’,’renovated’, ‘normed\_living (area)’, ‘age’, and ‘zipcode’ achieved the following results

