**ADTA 5120/CSCE 5300 Spring 2019**

**Warm-up Exercises**

**Due 2/3/18**

Prepared by Kuntal Patel

* 1. 1) (25 pts)
  2. The National Science Foundation’s Higher Education Research and Development Survey is the primary source of information on R&D expenditures at U.S. colleges and universities. The survey collects information on R&D expenditures by field of research and source of funds, as well as information on types of research and expenses and headcounts of R&D personnel. The survey is an annual census of institutions that expended at least $150,000 in separately accounted for R&D in the fiscal year.
  3. a) Give summary statistics (mean, median, SD, IQR) and boxplots on R&D expenditures across higher education institutions for each fiscal year.

1. steps:
2. 1) Gathered information from official website and given Excel data sheet.by analyzing footnotes in excel datasheet got some idea how to handle non-numerical data.

2)There are two options to handle non-numeric data 1) simply ignore it and not at all consider it in data analysis, and 2) Calculate or put approx. value from available data points from other years. The problem with this method is since expenditure will increase overall from 2008 to 2017.and if we average out later year values and impute that value in earlier year it may cause data discrepancy. Because in most cases expenditure value in year 2008 is less than year 2017.So option 1 seems better solution to handle non-numerical data. Surely in this method for each fiscal year we will have different number if data points.

1. 3) Excluded ‘na’, ’NA’ and ‘ne’ from the dataset.

na = not applicable (Either institution did not exist or not a separate entity)

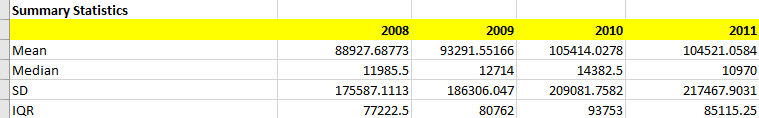
NA = not available; institution was not surveyed separately.

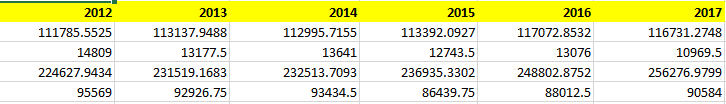
ne = not eligible; institution did not meet the criteria for inclusion.

For example, CUNY, Advanced Science Research Center is not having data from 2008 to 2014.In 2015 the expenditure is very less compared to overall dataset but still accountable. In 2016 again, it is not eligible to participate in survey and in 2017 it is a part of survey. So, for the proper data analysis purpose as data is not available from year 2008 to 2014 and 2016, I have not included this research center into calculation. But for year 2015 and 2017 have taken into consideration even if the values are not as significant as other data points.

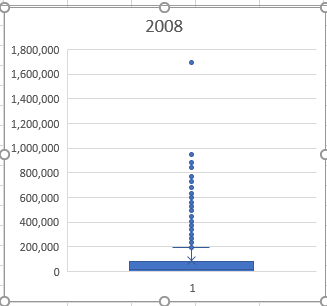
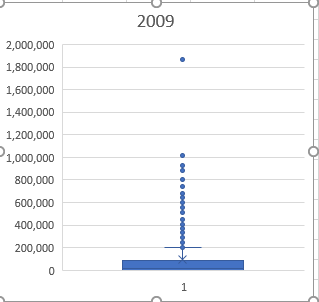
4) Now Data is clean, and I have different number of data points in each fiscal year because of removal of ‘na’,’ne’ and ‘NA’. Used Excel filter function, data analysis tool pack (Descriptive statistics) and quartile functions, Box and whisker plot to achieve clean and visualize data.

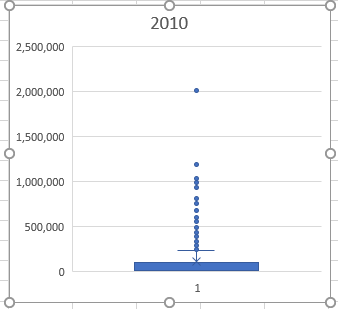
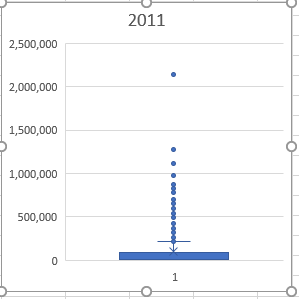
Results:

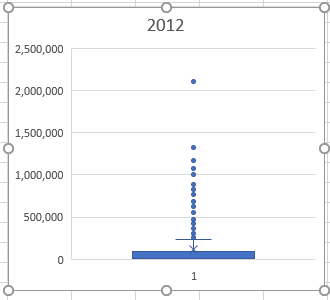
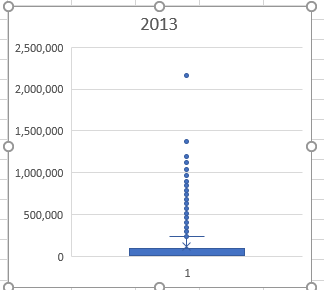


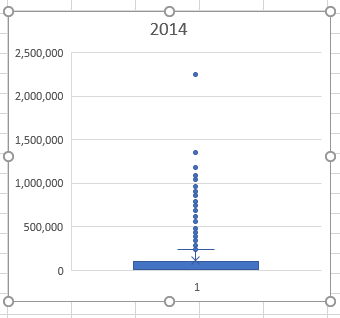
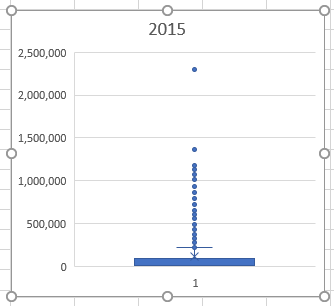


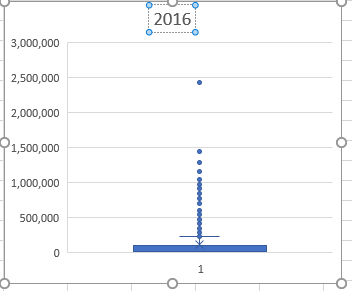
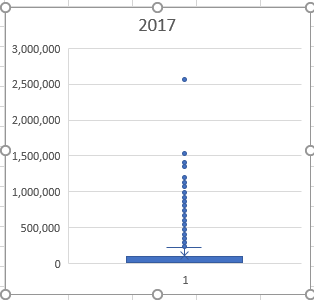
**Boxplots**

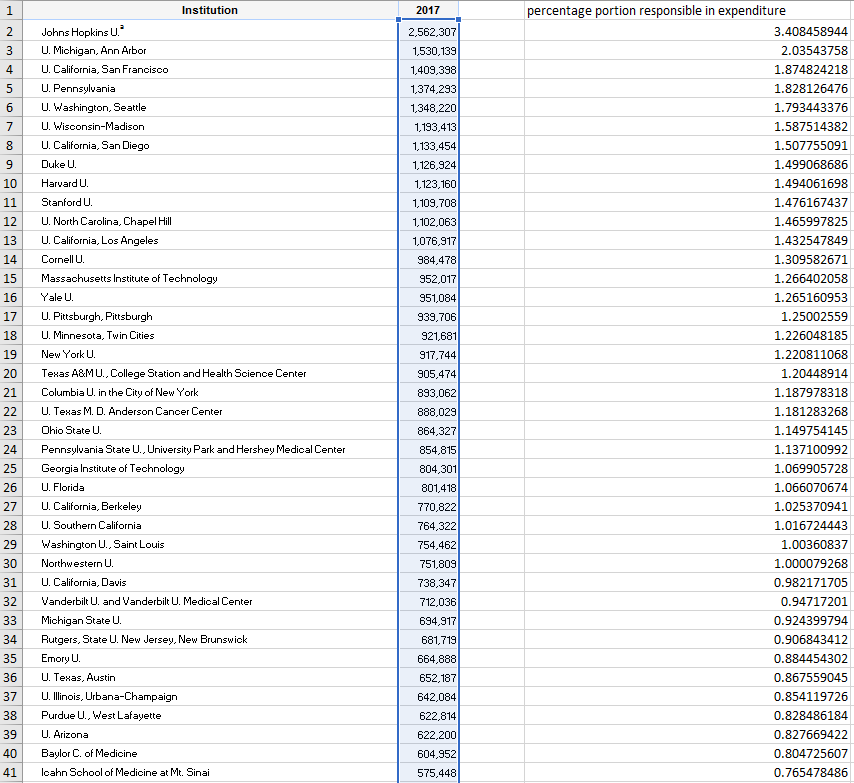
 

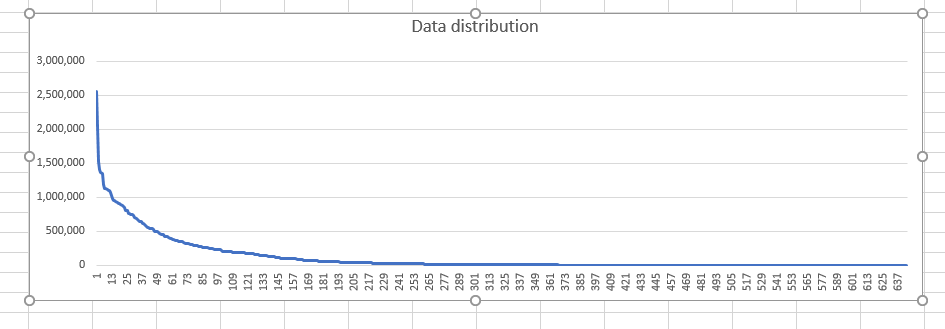
 

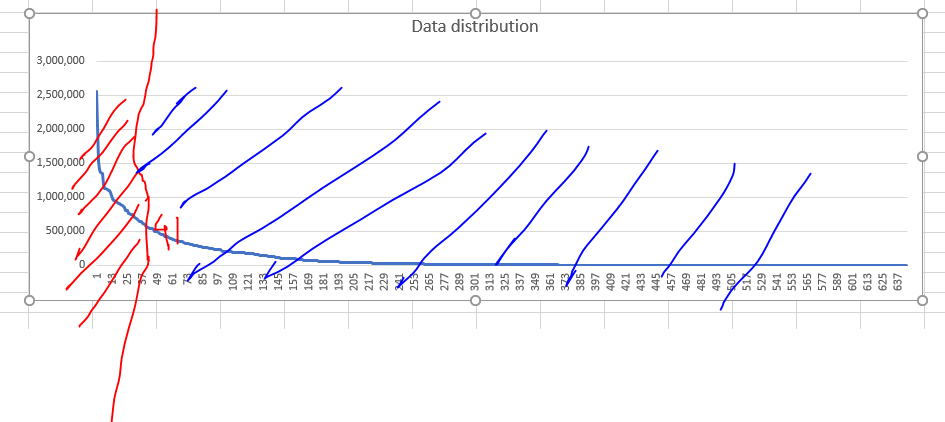
Analysis:

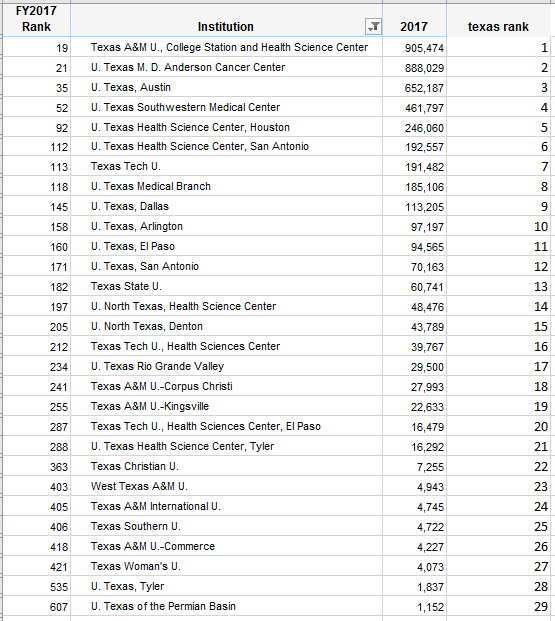
1. From summary statistics: mean and median value comparison it is evident for all fiscal year that values are not close, means data dispersion is not uniform. Average of all data points(mean) and most of the data(median) lies in different range. Standard deviation value is significant high i.e. most of the data point are far away from mean value.
2. From box plots can see number of outliers which indicates there are few universities are in outlier range (more than 1.5 times IQR) which has most of the expenditure in higher range. All other universities are below 500000 expenditure range.
   1. b) Analyze and describe the distribution of FY2017 R&D expenditures among institutions. For example, are R&D expenditures uniformly distributed across universities, or do a few universities account for the majority of expenditures? Between what values are the middle 50% of universities?
   2. In year 2017 R&D expenditures is not uniformly distributed. Almost 40 universities from total surveyed universities are accounted for 50% of expenditure.
   3. To analyze have sorted data of 2017 based on highest to lowest expenditure and took our percentage contribution in terms of expenditure.
   4.  
   5. From the line chart only, few (40) universities are responsible for expenditure, rest of the universities comes under very less values of expenditure.

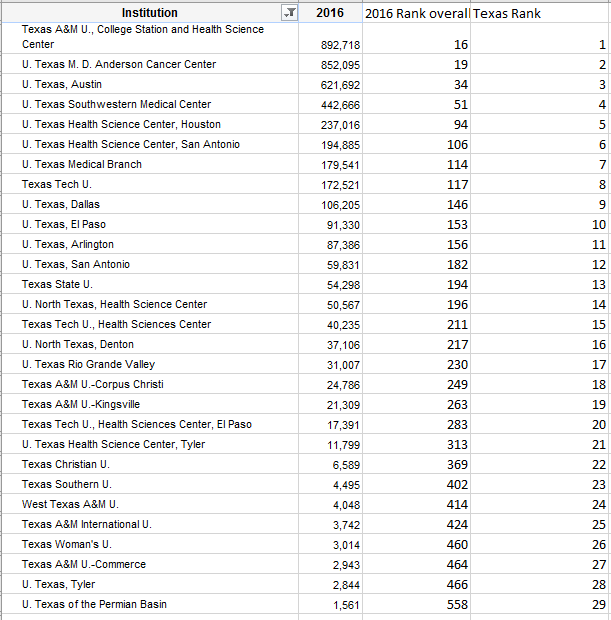
From IQR middle 50% universities lies between 93078.75 and 2494.75



Line chart in which almost 40 universities are contributing in 50% of expenditure.



* 1. c) Identify FY2016 and FY2017 rankings and percentiles of Texas universities (overall and among other Texas universities). Which Texas universities moved up in rankings and which moved down from FY2016 to FY2017 (overall and among other Texas universities).
     + For year 2017 ranking is given which is based on highest to lowest expenditure.
     + Filtered Texas universities and rank them among other universities. Texas has total 29 universities in survey. First column shows overall ranking of Texas universities.
  2. 
     + Same way calculated 2016 ranking

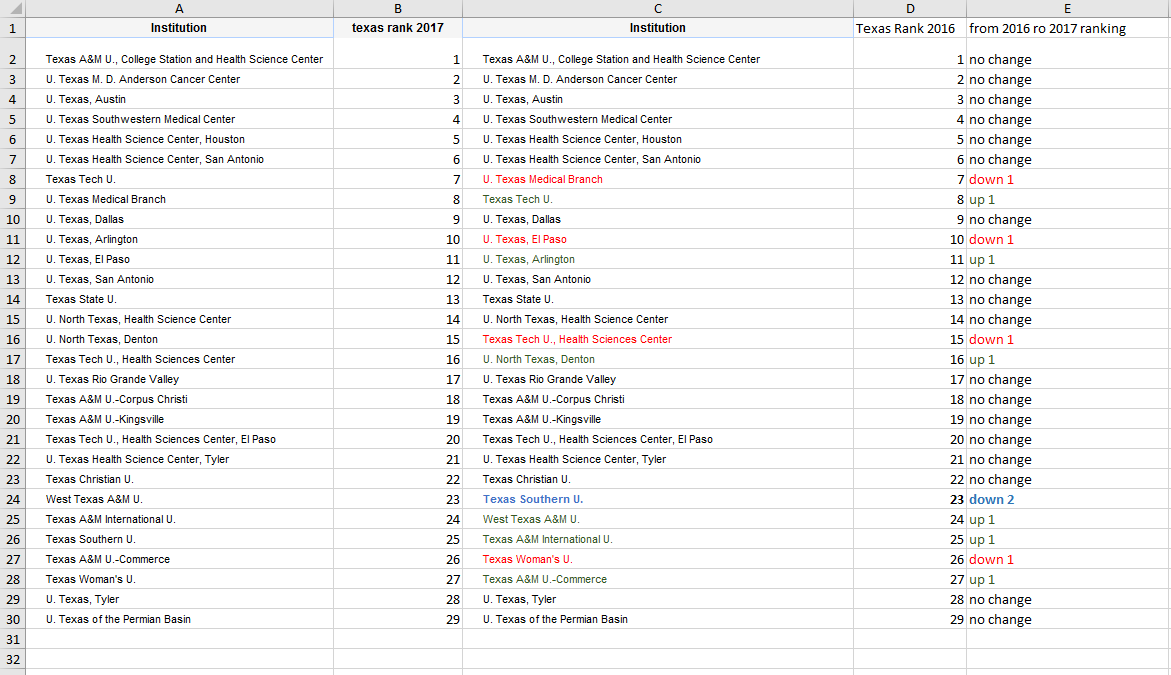


Comparision of ranking in year 2016 to 2017: U. Texas Medical Branch, U. Texas Medical Branch, Texas Tech U., Health Sciences Center, Texas Woman's U. universities lost 1 Rank from previous year.

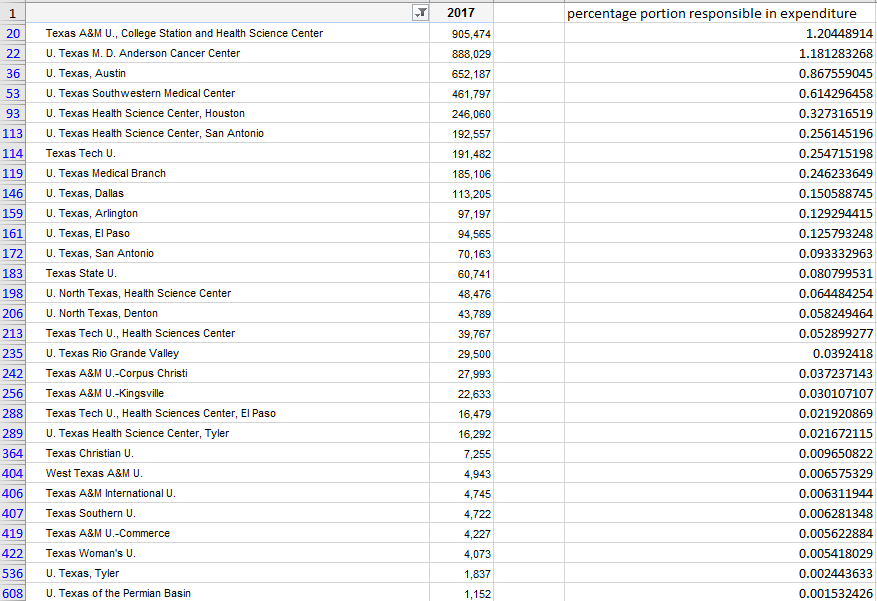
Texas Tech U., U. Texas, Arlington, U. North Texas, Denton, West Texas A&M U., West Texas A&M U., Texas A&M U.-Commerce universities gained 1 rank from previous year.

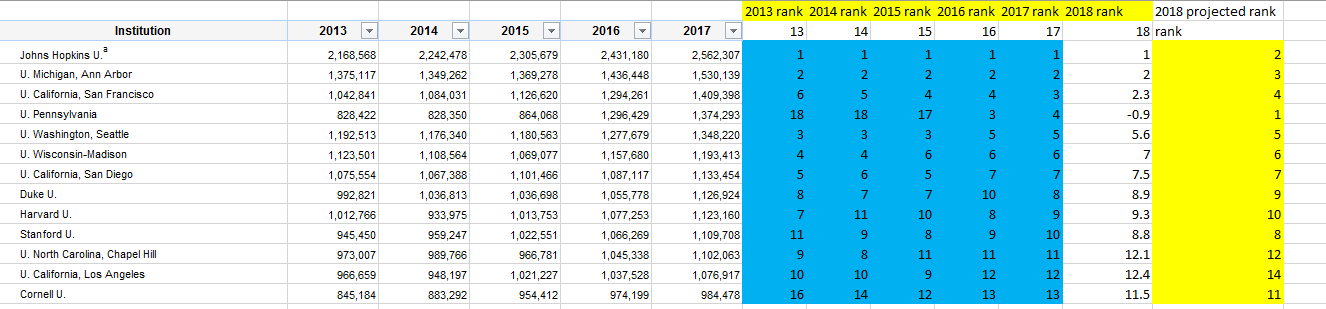
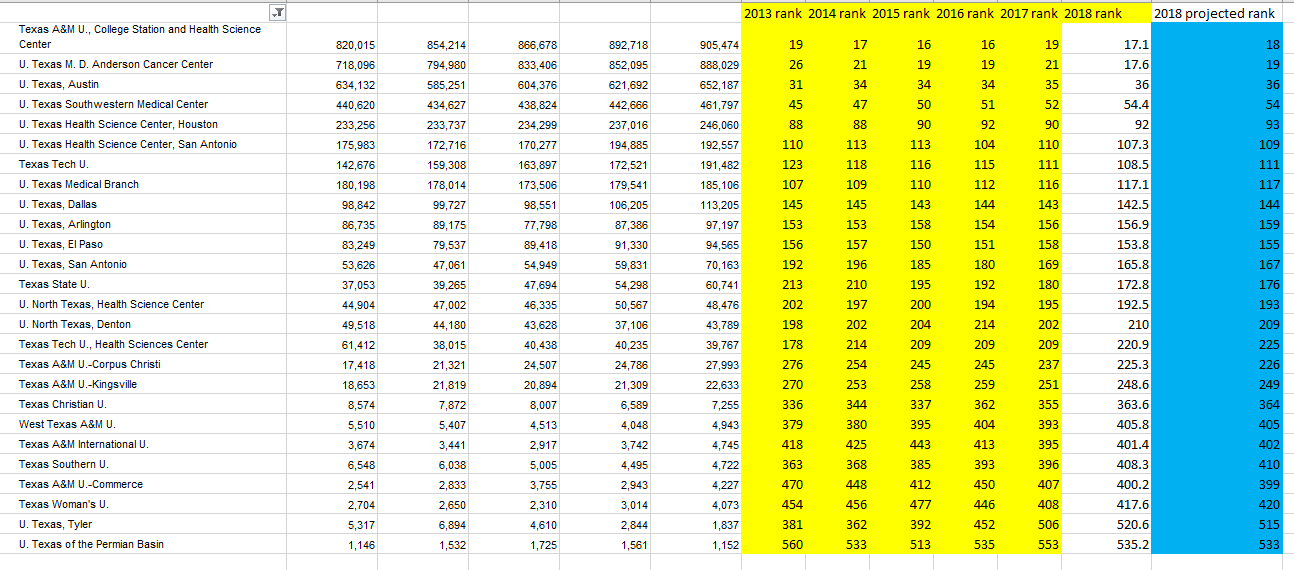
**Texas Southern U. lost 2 rank from its previous year.**

**Other universities are in same Rank for year 2016 and 2017.**

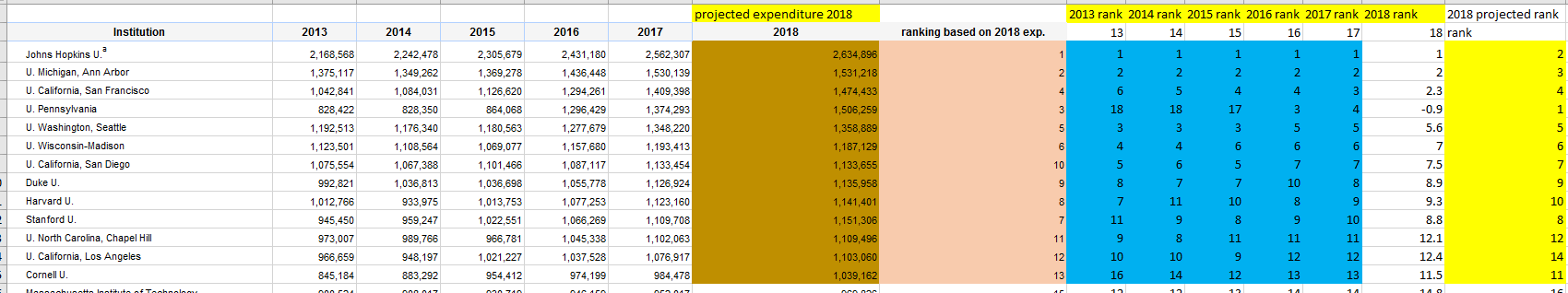


Percentage portion in overall expenditure in year 2017 of all Texas universities is around 6%.



* 1. 
  2. d) Based on the last five fiscal years, project R&D expenditures and overall rankings for Texas universities in FY2018. Do the projected expenditures imply any change in rankings?
  3. overall projected ranking in 2018: John Hopkins may come second in rank and University of Pennsylvania comes first in the Rank.
  4. In year 2013,2014 and 2015 John Hopkins is lower in Rank because of less expenditure, because Rank is calculated based on expenditure. Higher expenditure Universities are having ranks in lower number term.
  5. 
  6. In Texas universities
  7. 

1. Do the projected expenditures imply any change in rankings?
2. yes, but it depends on how we get the rank

* when projected 2018 expenditure just based on past 5-year data for each university ranks are different than when got the rank of 2018 year based on past five-year ranks.
* For example, John Hopkins will be in rank 1 position if we only consider past five-year expenditure of it. But when we consider all universities and rank it for 2018 it may come second because University of Pennsylvania is constantly jumping in ranks from 18 to 4 in five years.
  1. 

The data in *HERD FY17* gives research expenditures (dollars in thousands) for FY’s 2008 – 2017.

2) (25 pts) The *Wisconsin Breast Cancer* data consists of attribute information on fine-needle aspirated breast tissue from 569 women. Information on each sample includes

Column 1: Patient ID number

Column 2: Diagnosis (M = malignant, B = benign) 3-32)

Columns 3 – 32: Summary statistics on ten real-valued features computed from each cell nucleus in a sample

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

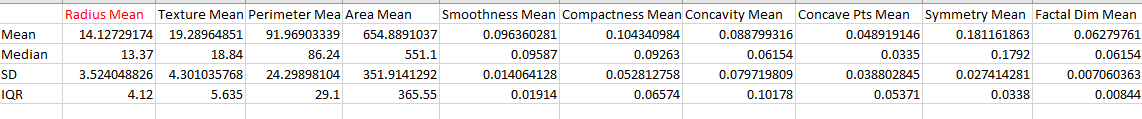
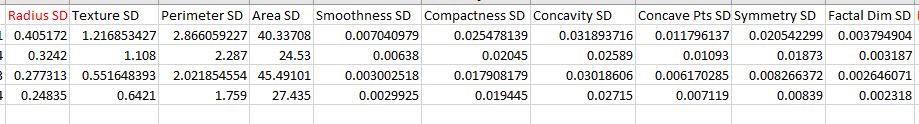
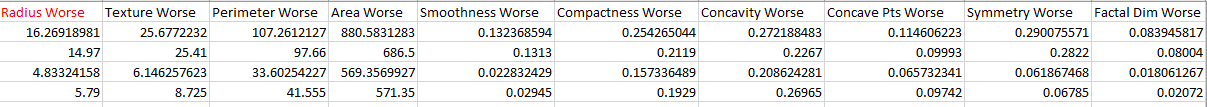
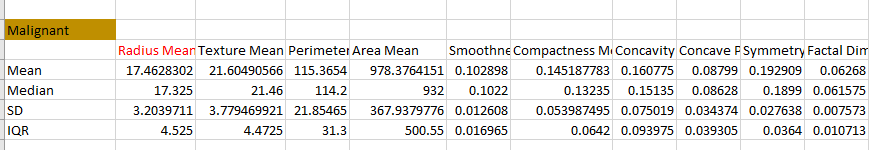
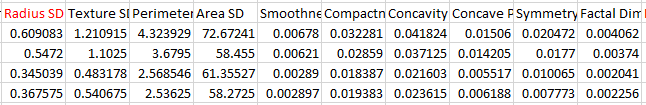
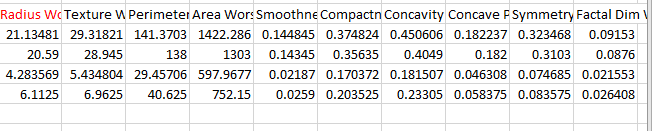
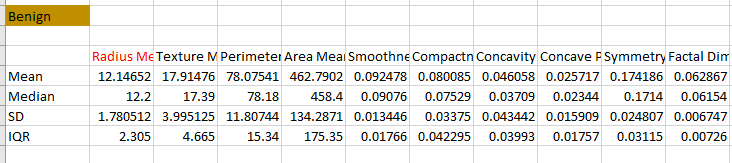
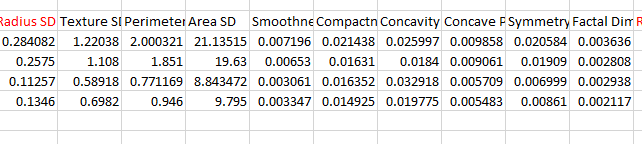
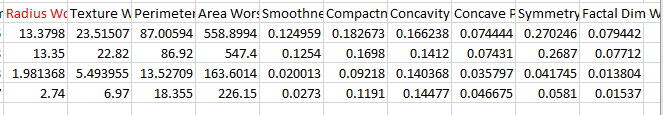
i) symmetry

j) fractal dimension ("coastline approximation")

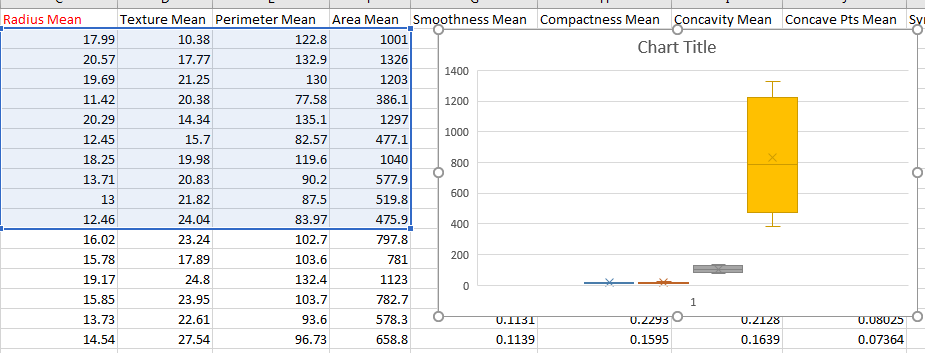
For each feature, the mean, standard deviation, and "worst" or largest (mean of the three largest values) were computed for each sample, resulting in 30 measurements. For example, column 3 is Mean Radius, column 13 is Radius SD, column 23 is Worst Radius.

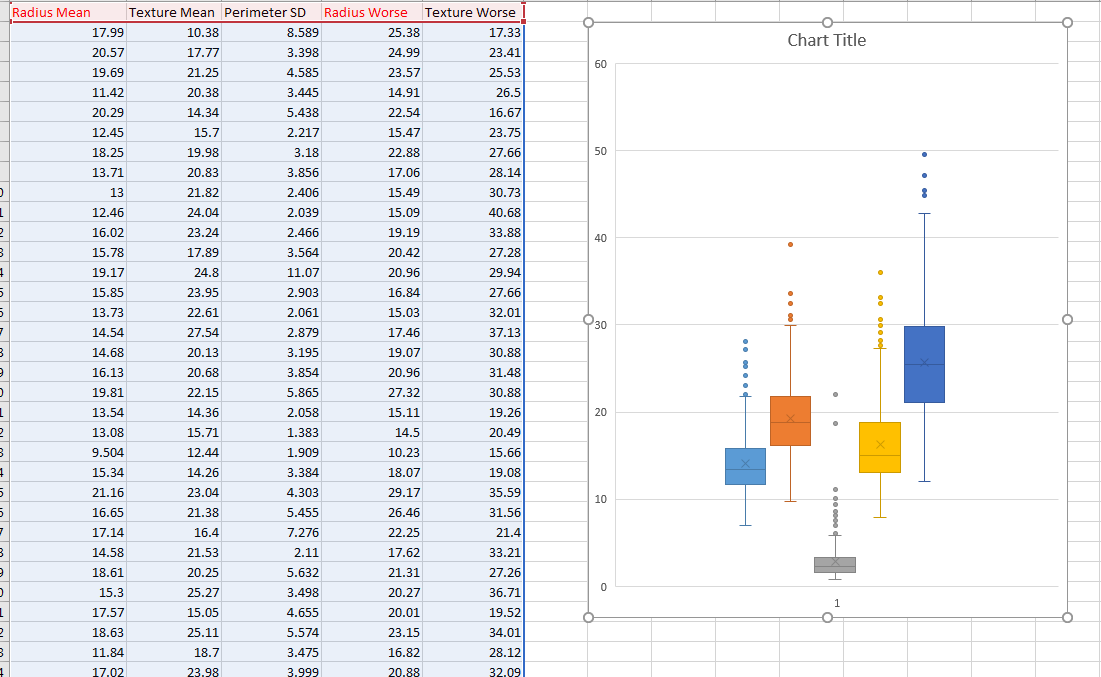
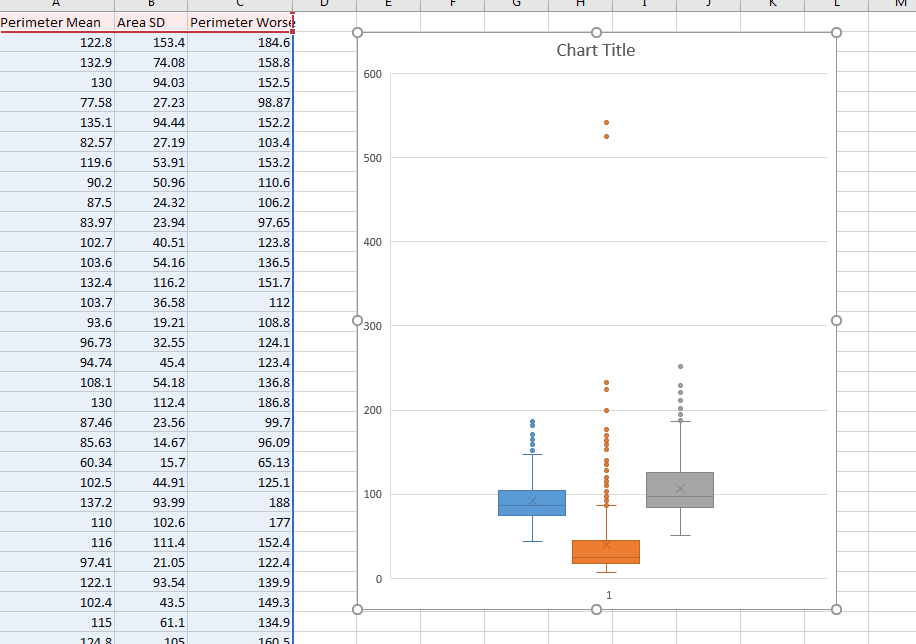
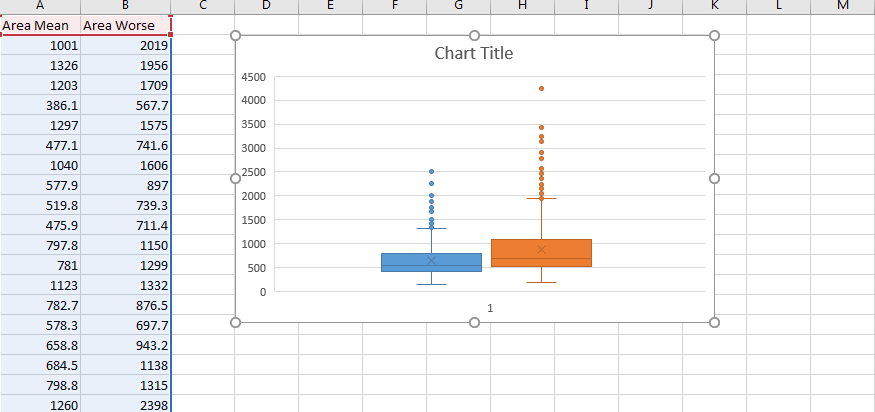
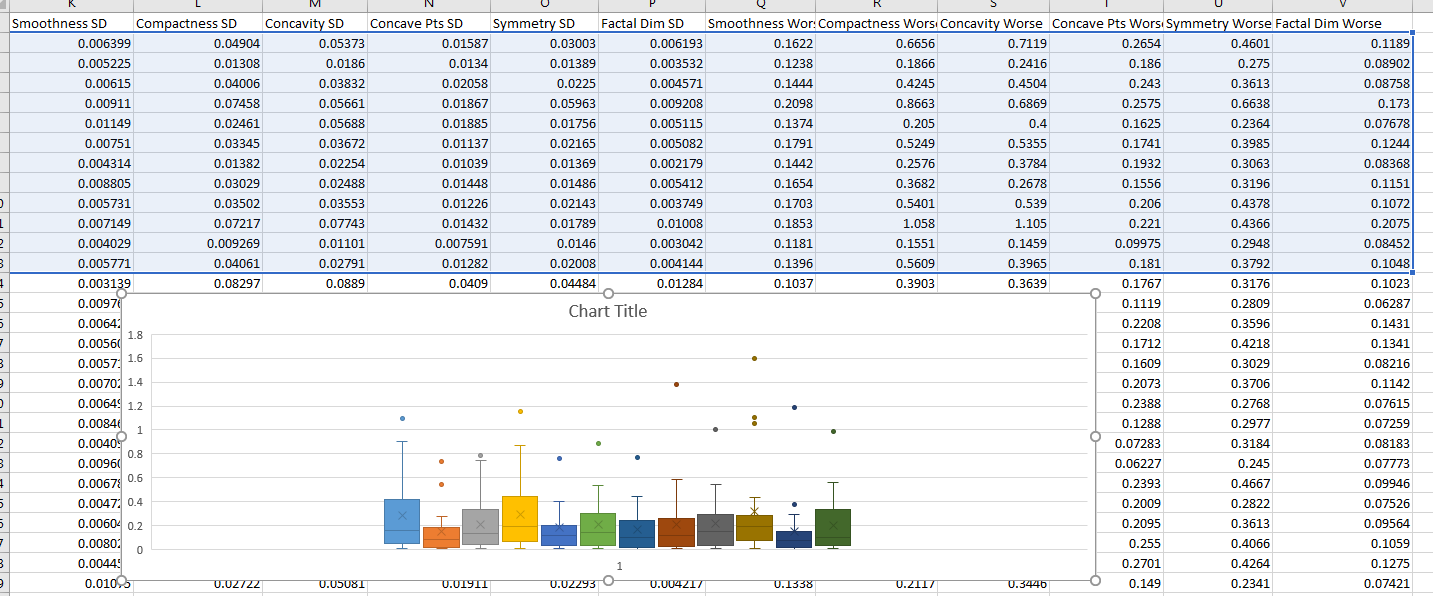
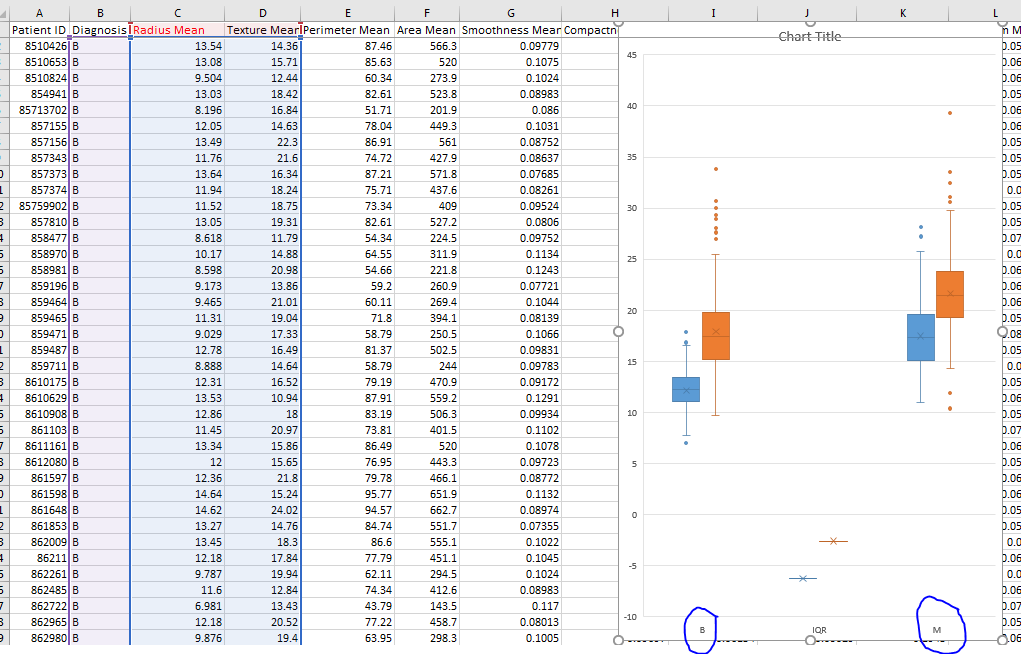
1. a) Give summary statistics (mean, median, SD, IQR) for each attribute feature – overall and with respect to diagnosis.

* Data is having all numerical values. Even some values are zero which may point very small value in the dataset and considered as a zero.
* Problem with this dataset is scale of columns which varies in significant way, hence when plot all the attributes on single boxplot for some attributes we could not able to see minute details and hence have to decide which attributes can be plotted parallelly.

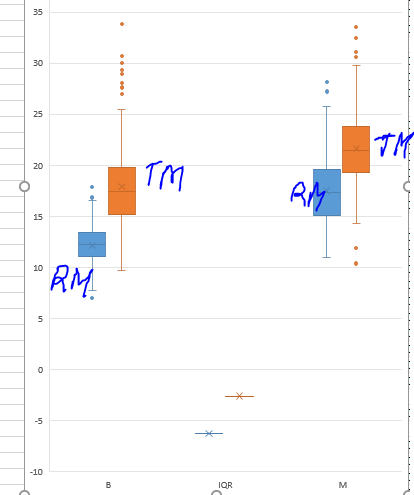
1. overall summary statistics
2. 
3. 
4. 
5. with respect to Diagnosis
6. w.r.t. Malignant :filtered data based on Diagnosis column
7. 
8. 
9. 
10. w.r.t Benign
11. 
12. 
13. 
14. b) Construct parallel boxplots for each attribute feature, comparing differences between diagnoses.

Scale of Radius mean, and Area mean is different and so from boxplot of radius mean, texture mean, perimeter mean along with area mean can not give enough information and comparison.



1. solution t this problem is plotting similar scale attributes together and then try to compare its results.
2. 
3. 
4. 
5. 
6. 
7. c) From the summary statistics and boxplots, suggest (and justify) guidelines for automating the diagnosis using the attribute features.
8. RM = Radius mean
9. TM = Texture mean
10. Analysis:
11. from below box plot if radius mean is below 14 it is more likely to have benign tumor and if it is above fifteen there is more likelihood of having tumor malignant. Same way for the texture mean, 20 is the threshold line.

This EDA analysis is useful to supplement any decision to diagnose tumor type.

1. 

**General Expectations for Assignment**

* • Clearly, concisely answer specific questions.
* • Justify claims with appropriate analysis and supporting data summaries and/or visualizations.
* • Provide an overview of the data, including a description of variables relevant to the analysis, variables not relevant (and why), and variable correlations/interactions.
* • Identify any limitations of the data in addressing problem objectives.
* • Use proper formatting, grammar, punctuation and clear exposition.
* • Turn in either MS Word, or PDF document.