README

> GitHub: https://github.com/GULS1377/NBA_Data_Analytics

1. Directory Structure

- PJ Phase2
  - PJ Phase1
    - add 21.salary_cleaned_newFormat.csv
  - Main.py
  - clean.py
  - salary_src.csv
  - salary_dst.xlsx
  - player_src.csv
  - player_dst.csv
  - player_career_src.csv
  - player_career_dst.xlsx
  - player_normalization_dst.csv
  - player_career_normalization_dst.csv
  - mean_median_std_player.txt
  - mean_median_std_player_career.txt
  - mean_median_std_salary.txt
  - lof.py
  - outlier_player_career_dataset.txt
  - plt.py
  - Clustering_ associationrules.py
  - Predictive_Analysis_Classification.py
  - Predictive_Models_Graphs.xlsx
  - t_test.py
  - README.md
  - README.pdf
  - Project Report.md
  - Project Report.pdf

2. Files Description

- PJ Phase1: the directory of our original PJ Phase 1.
- 21.salary_cleaned_newFormat.csv: change from PJ Phase1/9.salary_cleaned.xlsx, used for association rules part.
- Main.py: the main python file of the whole PJ Phase2.
- clean.py: main python for Basic Statistical Analysis and data cleaning insight.

- salary_src.csv: salary dataset from PJ Phase1.
- salary_dst.xlsx: salary dataset after cleaning in PJ Phase2.
- player_src.csv: play stats dataset from PJ Phase1.
- player_dst.csv: play stats dataset after cleaning in PJ Phase2.
- player_career_src.csv: play career stats dataset from PJ Phase1.
- player_career_dst.xlsx: play career stats dataset after binning in PJ Phase2.
- player_normalization_dst.csv: player stats dataset after normalization in PJ Phase2.
- player_career_normalization_dst.csv: player career stats dataset after normalization in PJ Phase2.
- mean_median_std_player.txt: output file for calculation of mean, median and std on player_stats dataset.
- mean_median_std_player_career.txt: output file for calculation of mean, median and std on player_career_stats dataset.
- mean_median_std_salary.txt: output file for calculation of mean, median and std on salary dataset.
- lof.py: main python file for Basic Statistical Analysis and data cleaning insight CS Extra Part.
- outlier_player_career_dataset.txt: output file including outliers' number.
- plt.py: main python for Histograms and Correlations part.
- Clustering_ associationrules.py: python file of Clustering and association rules.
- Predictive_Analysis_Classification.py: the python file for classification of Predictive Models Part.
- Predictive_Models_Graphs.xlsx: the excel file for some graphs used in Project Report of Predictive Models Classification Part.
- t_test.py: main python file of Hypothesis tests.
- README.md: the markdown file of README for PJ Phase2.
- README.pdf: the pdf file of README for PJ Phase2.
- Project Report.md: the markdown file of Project Report for PJ Phase2.
- Project Report.pdf: the pdf file of Project Report for PJ Phase2.

3. Codes Explanation

- clean.py

  > main python for Basic Statistical Analysis and data cleaning insight.

  - **Functions Description**

```
# remove '$' symbol
def salary_clean()

# the calculation of mean, median, and std for selected attributes of
data sets
def mean_median_std()

# normalize selected attributes on 2 dataset
# and export to 'xlsx' files for further analysis
# form 0, normalize player_career dataset
# form 1, normalize play_stats dataset
def normalization()

# if a player has more than 1 position, only use the 1st priority
position
def player_stats_position_remove()

# player_stats dataset clean
```

```
def player_stats_clean()

# equal-width binning on 'G' (Game Played) attribute into 5 class
def bin_data()
```

- lof.py

  > main python file for Basic Statistical Analysis and data cleaning insight CS Extra Part.

  - **Functions Description**

```
# using 3 different k-value and cross-check the result to determine the
outlier
def lof()

# run LOF to find outliers
def run_lof():
```

- plt.py

  > main python for Histograms and Correlations.

  - **Functions Description**

```
# Use a histogram to plot at least three of the variables
def hist():

# use scatter plot to show the correlation between attributes #
# to show the relation between season and player's mean salary
def scatter():
```

- Clustering_ associationrules.py

  > python file of Clustering and association rules.

  - **Functions Description**

```
# Function that for calculating the silhouette score for three methods
and draw the diagram
def plot()

#The function of PCA
def process_pca()

# The function of different clustering methods and measurements
def clustering()

# Function that contains apriori algorithm method
def print_apriori()

# The function runs association rule mining on a subset of
salary_of_players.csv
def association()

# More detailed comments are in the python program
```

- t_test.py

> main python file of Hypothesis tests.

- **Functions Description**

```
# execute the whole process
def execute():
```

- Predictive_Analysis_Classification.py

> the python file for classification of Predictive Models Part.

- **Functions Description**

```
# function that can initialize the new model
def initializeData()

# function that can get different normalized training data under different
norms
def getNorms()

# function that can calculate accuracy of different classifiers
def calculateAccuracy()

# function that can make non-numeric categorical data numeric
def categorical()
```

4. Group members and roles

Member	Roles
Lu Sun (ls1377)	Predictive Analysis Classification Part and integration of all documents for PJ Report and README.
Yijun Gan (yg270)	Clustering analysis and association rules analysis on subsets.
Shuyang Yu (sy614)	Predictive Analysis Hypothesis test Part.
Bo Zhou (bz166)	Basic Statistical Analysis and data cleaning insight, Additional Part for CS Students, Histograms and Correlations