**Project Report**

**Deliverables for this project**

1. **Team Name - Group 12**
2. **Team Members -**
   1. Mansi Parkhi
   2. Harshita Gupta
   3. Karthik Mandadi
   4. Harshitha Yata
   5. Lokesh Chowdary Vemula
   6. Sowri Madduru
   7. Shanmukh Kanneganti
3. **Introduction to Problem or Opportunity (include supporting domain information) -**

**Problem statement -**

Through this analysis, we are trying to perform NFL analysis and predict the type of play

**Supporting domain information -**

The National Football League (NFL) is the United States' largest and longest-running professional football league. The National Football League (NFL) is made up of 32 clubs divided into two conferences: the American Football Conference (AFC) and the National Football Conference (NFC) ( NFC ).

In a Football game, knowing whether the next play will be a run or a pass gives a huge edge to the defensive coordinator.

1. **Research Question(s) - Primarily, your research question should define the scope of your research, your data source and what you are seeking to discover.**

The purpose of this project is to explore NFL data from 2009-2017. In this project, We have loaded, explored, cleaned and modelled the data in order to anticipate the play type of the Football team named - 'Cleveland Browns' . Following that, We have checked if any adjustments can be made to fine-tune our best model.

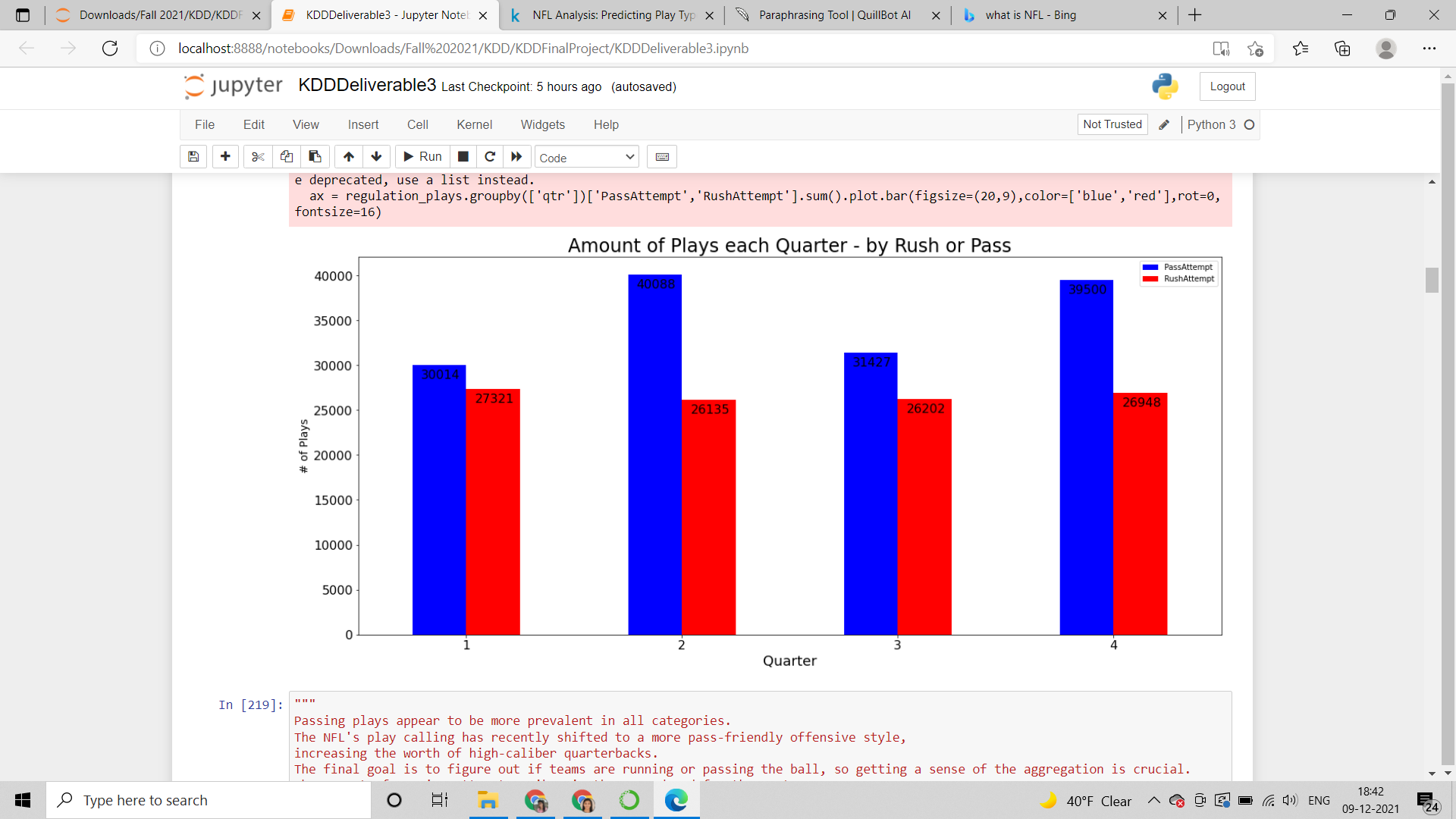
1. **Data Resources - you should have links and descriptions of where you are getting your data**

The data is obtained from - <https://www.kaggle.com/>. The creator of the data has used nflscrapR to obtain the data and also represents https://www.cmusportsanalytics.com/. nflscrapeR is an R package engineered to scrape NFL data by researchers Maksim Horowitz, Ron Yurko, and Sam Ventura.

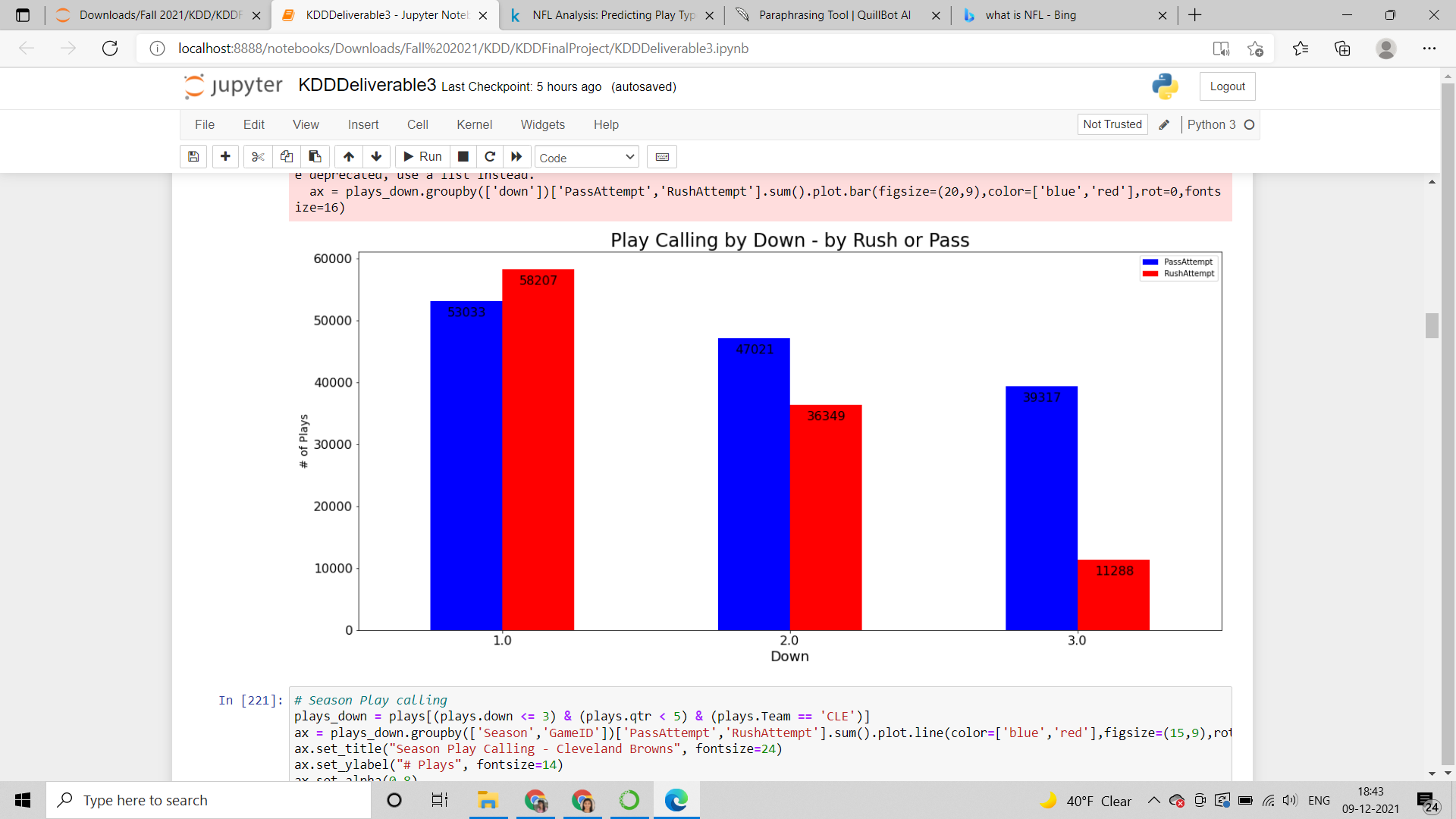
1. **Data Preprocessing** **-** We saw that the data was more than what we required, so we decided to shorten the scope of the columns. Also, First, we'll need to compile data on the team in possession of the ball's average yards gained on rushing and passing plays. This ensures that we can make decisions based on the teams' relative performance in that game. To account for 'huge plays' that can alter a coach's confidence in particular play calls, we used the Mean rather than the Median. Finding this value is critical to consider because it has a significant impact on a human decision. For data cleaning, we cleaned the float data and transferred it to integer.
2. **Data Understanding and Exploration** **- you should have some charts showing features of your data, it's distribution, etc. There should be some writing on what you learned from this and how it will inform what you do going forward.**

**Some charts which we have shown are -**

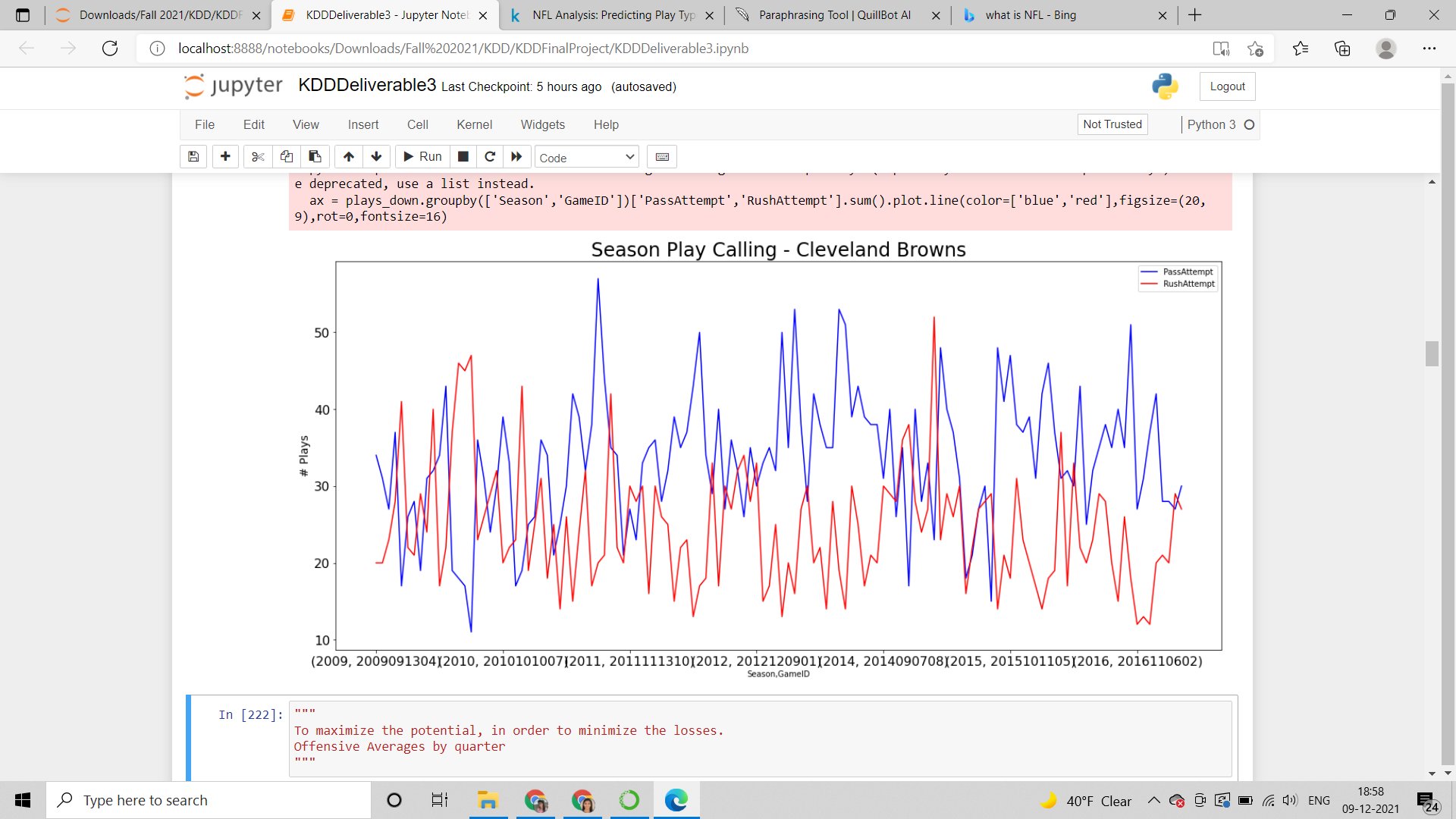
Amount of plays each quarter - Bar Graph

****

**Play call by down**

****

**Season Play Calling**

****

1. **Data Preparation for Modeling** **-**

We have segmented our dataset into training dataset and testing data, which made it easier for us to easily build our models and measure the accuracy of our model.

1. **Modeling -** We have built four models -
   1. Random Forest Classification
   2. K-Means Clustering
   3. Gaussian Naive Bayes
   4. GradientBoostingClassifier()
2. **Evaluation -** We measured the results using the AUC (Area Under Curve) method. We have used this method because it suits binary decision models well.
3. **Results -** GradientBoostingClassifier() is the best model
4. **Future Work -** We plan to take the analysis further by doing the comparison for different teams.
5. **Conclusion**
   1. What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

Ans. We saw that the data was more than what we required, so we decided to shorten the scope of the columns. We compiled data on the team in possession of the ball's average yards gained on rushing and passing plays to ensure that we can make decisions based on the teams' relative performance in that game. To account for 'huge plays' that can alter a coach's confidence in particular play calls, we used the Mean rather than the Median. Finding this value is critical to consider because it has a significant impact on a human decision. For data cleaning, we cleaned the float data and transferred it to integer.

* 1. Did you create any new additional features / variables?

Ans. We did not add any new features as of now, since we have already included many features in our analysis.

* 1. What was the process you used for evaluation? What was the best result?

Ans. We used 4 models for our evaluation - Random Forest Classification, K-Means Clustering, Gaussian Naive Bayes, GradientBoostingClassifier(). GradientBoostingClassifier() was the best.

* 1. What were the problems you faced? How did you solve them?

Ans. The data was very huge and we needed to shorten the scope of the data, and there was a lot of redundant data, which needed to be cleaned. We shortened the data and removed the float values from the data and took only integer values, which solved the problem.

* 1. What future work would you like to do?

Ans. As of now, we have done satisfactory analysis, in future, we plan to take the analysis further by doing the comparison for different teams.

* 1. Instructions for individuals that may want to use your work

Ans. There are no specific instructions.