PROJECT – STAT 355

Introduction-

The dataset I have chosen to use is a Kaggle dataset laid out by the National Football League as part of their Big Data Bowl initiative. The dataset has 7 separate data files which give a lot of information about the special teams plays which are one the three big facets of the game of football. I have used two of the 7 data files provided in the Kaggle competition to analyze data and make some valid inferences from it. The three datasets I have used are Plays.csv and PFFScoutingData.csv and Games.csv. This dataset is big compared to the ones we have used in class, so please take that into account. The link for the Kaggle page- https://www.kaggle.com/c/nfl-big-data-bowl-2022/data

Part-1 Exploring the dataset-

First to get a feel for our dataset we find their summary and see how the values for different columns vary and what type of value each column stores.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

Now, that we can see how each variable or column looks and we can use the mean and median to get an idea of their skewness.

Now, here’s a brief layman’s description of the dataset variables-

Play data

* gameId: Game identifier, unique (numeric)
* playId: Play identifier, not unique across games (numeric)
* playDescription: Description of play (text)
* quarter: Game quarter (numeric)
* down: Down (numeric)
* yardsToGo: Distance needed for a first down (numeric)
* possessionTeam: Team punting, placekicking or kicking off the ball (text)
* specialTeamsPlayType: Formation of play: Extra Point, Field Goal, Kickoff or Punt (text)
* specialTeamsResult: Special Teams outcome of play dependent on play type: Blocked Kick Attempt, Blocked Punt, Downed, Fair Catch, Kick Attempt Good, Kick Attempt No Good, Kickoff Team Recovery, Muffed, Non-Special Teams Result, Out of Bounds, Return or Touchback (text)
* kickerId: nflId of placekicker, punter or kickoff specialist on play (numeric)
* returnerId: nflId(s) of returner(s) on play if there was a special teams return. Multiple returners on a play are separated by a ; (text)
* kickBlockerId: nflId of blocker of kick on play if there was a blocked field goal or blocked punt (numeric)
* yardlineSide: 3-letter team code corresponding to line-of-scrimmage (text)
* yardlineNumber: Yard line at line-of-scrimmage (numeric)
* gameClock: Time on clock of play (MM:SS)
* penaltyCodes: [NFL categorization](https://operations.nfl.com/the-rules/2021-nfl-rulebook/#table-of-foul-codes) of the penalties that occurred on the play. A standard penalty code followed by a d means the penalty was on the defense. Multiple penalties on a play are separated by a ; (text)
* penaltyJerseyNumber: Jersey number and team code of the player committing each penalty. Multiple penalties on a play are separated by a ; (text)
* penaltyYards: yards gained by possessionTeam by penalty (numeric)
* preSnapHomeScore: Home score prior to the play (numeric)
* preSnapVisitorScore: Visiting team score prior to the play (numeric)
* passResult: Scrimmage outcome of the play if specialTeamsPlayResult is "Non-Special Teams Result" (C: Complete pass, I: Incomplete pass, S: Quarterback sack, IN: Intercepted pass, R: Scramble, ' ': Designed Rush, text)
* kickLength: Kick length in air of kickoff, field goal or punt (numeric)
* kickReturnYardage: Yards gained by return team if there was a return on a kickoff or punt (numeric)
* playResult: Net yards gained by the kicking team, including penalty yardage (numeric)
* absoluteYardlineNumber: Location of ball downfield in tracking data coordinates (numeric)

PFF scouting Data-

* gameId: Game identifier, unique (numeric)
* playId: Play identifier, not unique across games (numeric)
* snapDetail: On Punts, whether the snap was on target and if not, provides detail (H: High, L: Low, <: Left, >: Right, OK: Accurate Snap, text)
* operationTime: Timing from snap to kick on punt plays in seconds: (numeric)
* hangTime: Hangtime of player's punt or kickoff attempt in seconds. Timing is taken from impact with foot to impact with the ground or a player. (numeric)
* kickType: Kickoff or Punt Type (text).
  + Possible values for kickoff plays:
    - D: Deep - your normal deep kick with decent hang time
    - F: Flat - different than a Squib in that it will have some hang time and no roll but has a lower trajectory and hang time than a Deep kick off
    - K: Free Kick - Kick after a safety
    - O: Obvious Onside - score and situation dictates the need to regain possession. Also the hands team is on for the returning team
    - P: Pooch kick - high for hangtime but not a lot of distance - usually targeting an upman
    - Q: Squib - low-line drive kick that bounces or rolls considerably, with virtually no hang time
    - S: Surprise Onside - accounting for score and situation an onsides kick that the returning team doesn’t expect. Hands teams probably aren't on the field
    - B: Deep Direct OOB - Kickoff that is aimed deep (regular kickoff) that goes OOB directly (doesn't bounce)
  + Possible values for punt plays:
    - N: Normal - standard punt style
    - R: Rugby style punt
    - A: Nose down or Aussie-style punts
* kickDirectionIntended: Intended kick direction from the kicking team's perspective - based on how coverage unit sets up and other factors (L: Left, R: Right, C: Center, text).
* kickDirectionActual: Actual kick direction from the kicking team's perspective (L: Left, R: Right, C: Center, text).
* returnDirectionIntended: The return direction the punt return or kick off return unit is set up for from the return team's perspective (L: Left, R: Right, C: Center, text).
* returnDirectionActual: Actual return direction from the return team's perspective (L: Left, R: Right, C: Center, text).
* missedTacklers: Jersey number and team code of player(s) charged with a missed tackle on the play. It will be reasonable to assume that he should have brought down the ball carrier and failed to do so. This situation does not have to entail contact, but it most frequently does. Missed tackles on a QB by a pass rusher are also included here. Multiple missed tacklers on a play are separated by a ; (text).
* assistTacklers: Jersey number and team code of player(s) assisting on the tackle. Multiple assist tacklers on a play are separated by a ; (text).
* tacklers: Jersey number and team code of player making the tackle (text).
* kickoffReturnFormation: 3 digit code indicating the number of players in the Front Wall, Mid Wall and Back Wall (text).
* gunners: Jersey number and team code of player(s) lined up as gunner on punt unit. Multiple gunners on a play are separated by a ; (text).
* puntRushers: Jersey number and team code of player(s) on the punt return unit with "Punt Rush" role for actively trying to block the punt. Does not include players crossing the line of scrimmage to engage in punt coverage players in a "Hold Up" role. Multiple punt rushers on a play are separated by a ; (text).
* specialTeamsSafeties: Jersey number and team code for player(s) with "Safety" roles on kickoff coverage and field goal/extra point block units - and those not actively advancing towards the line of scrimmage on the punt return unit. Multiple special teams safeties on a play are separated by a ; (text).
* vises: Jersey number and team code for player(s) with a "Vise" role on the punt return unit. Multiple vises on a play are separated by a ; (text).
* kickContactType: Detail on how a punt was fielded, or what happened when it wasn't fielded (text).
  + Possible values:
    - BB: Bounced Backwards
    - BC: Bobbled Catch from Air
    - BF: Bounced Forwards
    - BOG: Bobbled on Ground
    - CC: Clean Catch from Air
    - CFFG: Clean Field From Ground
    - DEZ: Direct to Endzone
    - ICC: Incidental Coverage Team Contact
    - KTB: Kick Team Knocked Back
    - KTC: Kick Team Catch
    - KTF: Kick Team Knocked Forward
    - MBC: Muffed by Contact with Non-Designated Returner
    - MBDR: Muffed by Designated Returner
    - OOB: Directly Out Of Bounds

## Game data

* gameId: Game identifier, unique (numeric)
* season: Season of game
* week: Week of game
* gameDate: Game Date (time, mm/dd/yyyy)
* gameTimeEastern: Start time of game (time, HH:MM:SS, EST)
* homeTeamAbbr: Home team three-letter code (text)
* visitorTeamAbbr: Visiting team three-letter code (text)

Now, that we have some context about the large amount of variables we have at our disposal we can perform EDA on some select variables to get some more insight on the dataset-

First we look at the plays dataset to see how the different types of plays stack up to each other-

Text

Description automatically generated

Chart

Description automatically generated

Or a better looking and more understandable graph with proper labels which was obtained using the code in the R file-

Chart, bar chart

Description automatically generated

This shows that kickoff and punt are two most common play types. So, the questions I will perform hypotheses on will involve kickoffs and punts.

Now, to examine the PFF scouting dataset, we will look at how are kick types and punt types distributed see if there is a need to group them into different dataset or not. The code for generating these two tables is also in the R file-

Chart, histogram

Description automatically generated

This graph takes all kickoff kicks from the PFF scouting dataset.

From, this graph we can see that the Deep kick style is the most frequent kickoff style which makes sense as the deep kick into the Endzone is the target when kicking off for every kicker in the national league and it also presents the safest option for the kicking team, as it moves gets the ball as far as back away from the kicking teams endzone as possible.

A screenshot of a computer

Description automatically generated with medium confidence

This graph filters out to so that we only have all punt style kicks from the PFF dataset.

Now, we can see here that there is a more even distribution here, so we can maybe use grouping here, and see how different type of punt styles affect the outcomes.

We can also have a preliminary look at the grouping for punts, see if punt style affects kick hang time which basically means how long the kick stayed airborne. In general, punts generally have longer hangtimes as they are kicked from a snap instead from a tee like a kickoff. Also, we can get an insight on if a certain punt style has more hangtime than other punt styles.

Chart, box and whisker chart

Description automatically generated

So, we can see that of the two groups “N” and “A” that we have significant amount of data on, we can see that “A” style or Aussie style kicks tends to have more hangtime as compared to “N” or nose style kickoffs. We can also see that “R” style could be a technique which is specifically used to get lower hang-time possibly for shorted length kicks.

Now, here I would like to lay out some guidelines and points about the data here-

As the dataset here has a lot of variables, I will filter out and remove some text-type variables that do not have a lot of use when I answer certain questions.

Part-2

Laying out the Hypotheses-

Question-1 – Is returning a kickoff by the receiving team, worth it if the kick makes the end-zone of the receiving team?

Background information – Some important information about the kickoffs. When the kicking team kicks off, if the ball is caught in the end-zone by the receiving team, the receiving player can take a knee in the endzone and take what is called a touchback. This touchback rule allows the receiving team to start their drive off at their 25 yard line automatically. This touchback rule was introduced to decrease injuries on kickoffs. Now, the receiving team always has an option to return a kick regardless of the kick catch point, the advantage the receiving team gets from returning a kick that reaches the endzone is that they can attempt to return the kick more than 25 yards and gain an advantage over the touchback rule.

Setting the question – First we take a look at the data and see how kickoffs stack up. The play result variable gives out how many yards the kicking team gained on the play, so the less the play result value the better it is for the receiving team.

Now, if the receiving team takes a touchback then the play result value will automatically be 40 yards, as the kick is taken form the 35 yard line of the kicking team.

dataKickoffs = dataPlays[dataPlays$specialTeamsPlayType == "Kickoff",]

KReturn <- dataKickoffs[dataKickoffs$playResult!=40,]

KTouchback <- dataKickoffs[dataKickoffs$playResult==40,]

KReturnD <- KReturn[KReturn$kickLength > 65,]

Now, these four lines of code divide the DataPlays dataset into relevant parts. KReturn gives the kickoffs that were not touchbacks and KTouchback gives the kickoffs that were touchbacks. Now, we need to make one more modification in our dataset, as we are assessing returning a kick is worth it we can only take kicks into consideration that reach the endzone, so we only take kicks greater than 65 yards.

Chart, histogram

Description automatically generated

Now, the above histogram is a distribution of the kicks that reached the endzone and were returned. So, we can see the histogram is left skewed with a few outliers on the left side that show that a small amount of kick Returns were wildly successful, but overall a lot of kick were returned mainly between 5-40 yards.

Chart, box and whisker chart

Description automatically generated

Now, the boxplot also gives the same result as the histogram showing that the distribution is left skewed.

Now, if we take a mean of the returned kicks set, we can see that mean is 40.83225 which shows that on an average kicking teams lose a yard compared touchbacks when they return a kick.

Now, we also take a look at the returned kicks data set without the outliers to kind of a better sense, of how the distribution looks if we remove outliers or the one-off miracle plays that happen rarely.

The mean for the 1/5th trimmed dataset, is 42.3 which shows that on an average returning lose out on 2.3 yards when they return a kick instead of taking a touchback.

Chart, histogram

Description automatically generated

Now, after removing the outlier values and keeping the values within the inter-quartile range we get a new distribution which looks close to a normal distribution albeit slightly being a little left skewed. So, now that we have a close to normal distribution we can go out and perform tests.

Now, as our goal is to see if it is worth it to return a kick that can be taken as a touchback, I chose to use a hypothesis test, to test that out.

Here, my H0- or initial hypothesis is that the population mean for playResult is 40 which is the what the playResult value would be if all endzone kicks we taken as touchbacks should be on average equal to the kicks that were returned.

Now, my H1- or secondary hypothesis, is that the on an average the sample mean of all playResult values for kicks that were returned is greater than the assumed population mean value of 40 yards.

Here, when we run the greater than hypothesis t-test we get a small p-value less than 0.05 which shows that we can reject the initial hypothesis and say that on an average for a given sample, if a kick is returned its playResult value would be greater than 40 yards.

This shows that by returning the kicks that can be taken as touchbacks, teams are costing themselves valuable yards.

As a side note I also performed the t-test on the dataset with outliers and still got a low p-value which shows that returning kicks that can be taken as touchbacks can really hurt the receiving teams.

Question-2- Going for it on 4th down in punting situations-

Background information- NFL teams have been notorious for punting and not going for 4th downs for a long time. However, over the last few years, analytics have become a big part of the teams gameplan. So, we are going to see if teams have changed their habits season by season by seeing how the average yards to go for a punt has changed over the 3 years the dataset spans. For this question we are using the Plays dataset and the Games dataset, and we will combine them together in R.

Inference- if the average yards to go has increased year by year then this means that teams have been going for it on fourth down and not punting.

A screenshot of a computer

Description automatically generated with medium confidence

Now, after combining the two datasets, we look at the summary to understand the combined dataset.

Now, group our dataset into the three given season 2018, 2019 and 2020 and see their boxplots, to see how their variance varies over the three years, and if we can use ANOVA here.

Chart, box and whisker chart

Description automatically generated

Now, the above boxplot is a zoomed version of the original boxplot to give us a better idea.

From the two boxplots we can say that there is not much change between 2018-2019 if any but yards to go changed by a fair margin in 2019-2020.

So, our initial hypothesis or H0 is that mean(2018) = mean(2019) = mean(2020).

H1: atleast one mean differs.

Now, we are going to run the ANOVA analysis and see the results-

Text, letter

Description automatically generated

We can see that the p-value here is 0.02 which Is not very low but still less than the threshold value of 0.05.

So, we can reject the H0 hypothesis and accept H1 that at least one mean was different.

To analyze which mean was different we can use this test-

Text

Description automatically generated

We can see that there was not much variation in coaching attitudes between 2018-2019 but there is a significant change from 2019 to 2020.

Chart, line chart

Description automatically generated

Now, the QQ-plot seems to be iffy but we can still make an assumption that data comes from a normal distribution, as the line after a certain quantile appears to be straight enough.

This shows that the teams changed their thinking in going for it on fourth downs in punting territory.

Graphical user interface, text, application

Description automatically generated

The above code is just a sidenote to show how teams have changed their thinking on going for it. It is a big change if you compare year 2018 and 2020.

Question-3 – does punt direction affect kick length?

Background information- This time our question is simple enough to understand we are checking if a certain punt direction results in longer/shorter kicks than other ones? I chose this particular question because we have a pretty even distribution of punts here with about even distribution between center and non-center punts.

For reference- Center-C-1, Left-L-2, Right-R-1

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

This is how I converted the Kick directions into numeric type so that we can analyze it further.

Chart, box and whisker chart

Description automatically generated

Now, we have a boxplot here that shows that there is significant difference in medians of all 3 kick direction types. Their variances are also quite large especially for the center kick-type which is to be expected as it has the largest number of values in our dataset.

Now, we lay out our initial hypothesis, H0: mean(center), mean(right), mean(left) are all equal

H1: atleast one of mean(center), mean(right), mean(left) is different.

Now, we can run the anova-

Graphical user interface, text, application, email

Description automatically generated

From the results of the ANOVA we get a p-value less than the threshold 0.05 which shows that atleast one mean is different.

To further analyze the means we use the pariwise Bonferroni test. Here, we can see that there is not much variation between sample means between 1 and 2, but there is a significant difference in 1 and 3 and between 2 and 3 means.

Now, we can have a look at our means to support our ANOVA results-

Graphical user interface, text, application

Description automatically generated

So, we can say with certainty that average of right kicks generally tend to be the longest when compared to center or left.

To see check our assumptions we can take a look at our qq-plot for the anova values-

Chart, line chart

Description automatically generated

The plot is pretty much like a straight line with some iffy parts at the tail ends but that is to be expected as it is a pretty large dataset, so, we can say that the distribution is normal and we can say our use of ANOVA is valid.