DATA 606 Data Project Proposal

William Jasmine

Data Preparation

The data I will be using for this project comes from Kaggle, and contains data pertaining to about 22k proton-proton collisions that were observed in CERN's large hadron collider (LHC) while it was running in 2010. The data is stored on Github, and loaded in the cell below:

```
# load data
link <- getURL("https://raw.githubusercontent.com/williamzjasmine/CUNY_SPS_DS/master/DATA_606/Final_Pro
df <- read_csv(link, na=c("", "NA"))
```

Research question

In 2010 when these collisions took place, the Higgs Boson particle existed only as part of theoretical physics models. In 2012, its existence was famously proven experimentally thanks to the work of the thousands of physicists working at the CERN laboratory. One of the ways that they were able to prove its existence was by analyzing the "b-jets" created after particle collisions. A jet in particle physics refers to the similarly directed stream of particles that follow a quark or gluon that has flown off on its own as the result of a particle collision. Essentially, smashing two subatomic particles (protons, in this case) results in sending the elementary components of those particles (quarks, in this case) off in different directions. Thanks to quantum mechanics, these ultra-fast flying quarks pull new quarks and gluon pairs out of the surrounding vacuum and sends those particles flying in roughly the same direction, creating what looks like a jet of particles. b-jets are those jets that originate from a type of quark known as a bottom quark. b-jets were of particular interest to the scientists are CERN due to the fact that there are a number of larger particles that end up decaying into bottom quarks. In 2010 the Higgs Boson was hypothesized to be one of these larger particles, hence the reason they are counted and included as their own column in this data set `nBjets`.

As such, we can use this data set to answer the research question: *what are the conditions required for a particle collision to result in a b jet?*

Though we do now do know that this Higgs Boson exists, there are still a number of other hypothesized particles that have similar decay patterns resulting in bottom quarks. As such, there is still relevance in asking this research question.

Cases

What are the cases, and how many are there?

```
nrow(df)
```

```
## [1] 21726
```

As shown in the output above, there are 21,726 cases (particle collions) in this data set.

Data collection

Describe the method of data collection.

This data was collected in the CERN laboratory using a myriad of advanced detectors and digital sensors. The CERN website offers much greater detail into how these tools work to capture information about particle collisions.

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The link to the data can be found on Kaggle but also directly from the CERN website.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable in this case is the number of b-jets created by the particle collision, stored in the data set as `nBjets`. It is a quantitative, continuous variable.

Independent Variable(s)

The table below includes a list of the independent variables, as well as their descriptions:

Column Name	Description
MR	First razor kinematic variable, the MR variable is an estimate of an overall mass scale, which in the limit of massless decay products equals the mass of the heavy parent particle.
Rsq	Second razor kinematic variable, the Rsq variable is the square of the ratio R, which quantifies the flow of energy in the plane perpendicular to the beam and the partitioning of momentum between visible and invisible particles.
E1	Energy of the leading megajet.
Px1	x-component of the momentum of the leading megajet.
Py1	y-component of the momentum of the leading megajet.
Pz1	z-component of the momentum of the leading megajet.
E2	Energy of the subleading megajet.
Px2	x-component of the momentum of the subleading megajet.
Py2	y-component of the momentum of the subleading megajet.

Column Name	Description
Pz2	z-component of the momentum of the subleading megajet.
HT	The scalar sum of the transverse momentum of the jets.
MET	The magnitude of the vector sum of the transverse energy of the particles in the event.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

The following cell uses the summary function to get some of the descriptive statistics for all variables:

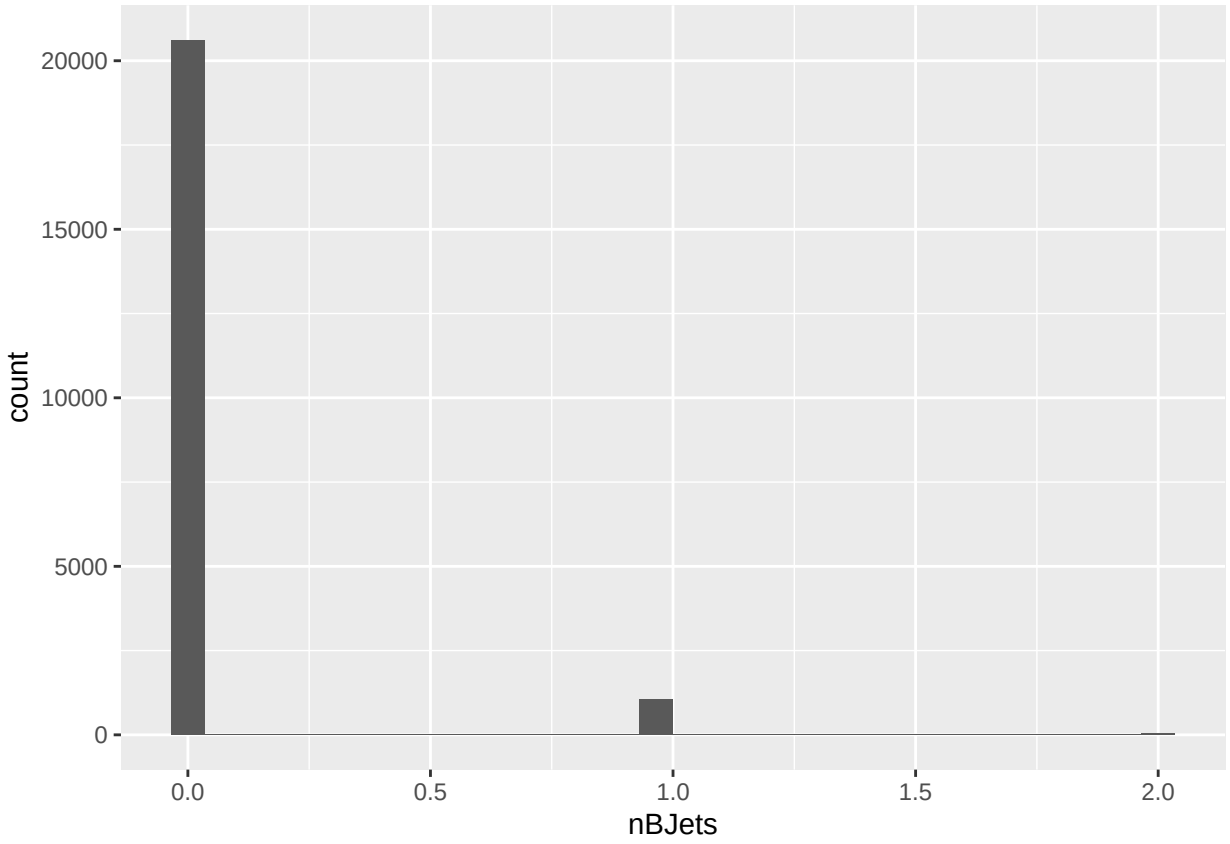
```
summary(df)
```

```
##       Run              Lumi            Event                 MR
##  Min.   :147926   Min.   : 388.0   Min.   :3.023e+08   Min.   :  30.0
##  1st Qu.:148029   1st Qu.: 463.0   1st Qu.:4.976e+08   1st Qu.: 229.8
##  Median :149181   Median : 986.0   Median :9.709e+08   Median : 292.9
##  Mean   :148666   Mean   : 900.7   Mean   :8.636e+08   Mean   : 352.0
##  3rd Qu.:149181   3rd Qu.: 999.0   3rd Qu.:9.845e+08   3rd Qu.: 406.9
##  Max.   :149181   Max.   :1804.0   Max.   :1.704e+09   Max.   :2433.8
##       Rsq                 E1               Px1                 Py1
##  Min.   :0.0000089   Min.   :  44.95   Min.   :-543.8210   Min.   :-648.3850
##  1st Qu.:0.0079072   1st Qu.: 143.53   1st Qu.: -78.4730   1st Qu.: -75.7700
##  Median :0.0168165   Median : 212.06   Median :  -0.3105   Median :   1.2387
##  Mean   :0.0232533   Mean   : 297.18   Mean   :   0.2833   Mean   :   0.7642
##  3rd Qu.:0.0316123   3rd Qu.: 374.54   3rd Qu.:  78.5625   3rd Qu.:  77.7381
##  Max.   :0.7636950   Max.   :2101.58   Max.   : 722.2910   Max.   : 470.2340
##       Pz1                 E2               Px2                 Py2
##  Min.   :-2022.310   Min.   :  42.05   Min.   :-700.1120   Min.   :-459.8010
##  1st Qu.: -151.348   1st Qu.: 126.92   1st Qu.: -63.3332   1st Qu.: -62.7064
##  Median :   -5.478   Median : 204.14   Median :  -0.5482   Median :  -1.9005
##  Mean   :   -8.523   Mean   : 277.41   Mean   :  -0.3994   Mean   :  -0.9099
##  3rd Qu.:  135.345   3rd Qu.: 366.71   3rd Qu.:  62.8520   3rd Qu.:  61.1116
##  Max.   : 2061.890   Max.   :1843.36   Max.   : 405.3260   Max.   : 635.7340
##       Pz2                 HT               MET                nJets
##  Min.   :-1647.600   Min.   : 120.9   Min.   :  0.1004   Min.   :2.000
##  1st Qu.: -154.232   1st Qu.: 193.3   1st Qu.:  8.6268   1st Qu.:2.000
##  Median :   -1.803   Median : 223.7   Median : 14.0350   Median :2.000
##  Mean   :   -1.915   Mean   : 242.3   Mean   : 16.0054   Mean   :2.436
##  3rd Qu.:  151.168   3rd Qu.: 269.2   3rd Qu.: 21.0910   3rd Qu.:3.000
##  Max.   : 1830.370   Max.   :1462.6   Max.   :423.1440   Max.   :7.000
##      nBJets
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.05367
##  3rd Qu.:0.00000
##  Max.   :2.00000
```

The histogram below gives the distribution of the number of b-jets formed from collisions:

```
ggplot(data = df, aes(x = nBJets)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Given the small number of collisions that actually do result in b-jets and the fact that even fewer have more than 1 b-jet, it might be better to instead create a new dependent, categorical variable `bjets` such that:

- `bjets = TRUE` if the collision did result in a b-jet.
- `bjets = FALSE` if the collision did not result in a b-jet.

This new variable is created below:

```
df <- df %>%
  mutate(bjets = ifelse(nBJets > 0, TRUE, FALSE))

table(df$bjets)
```

```
##
## FALSE  TRUE
## 20615  1111
```

The output above gives the specifics concerning the number of collisions that result in b-jets, and those that do not.

```
n = nrow(df)
p = unname(table(df$bjets))[2] / n

n * p >= 10
```

```
## [1] TRUE
```

```
n * (1 - p) >= 10
```

```
## [1] TRUE
```

The output above proves that if p is the probability of those collisions that resulted in a b-jet then np and $n(1-p)$ are each greater than 10, a requirement to conduct inference on a categorical variable.