Project Part 3

Research Question and Data Explanation

Based on the primary source data found on workplace injures, The research question most relevant to answer would be:

Is there any industry or subset of industry that is most prone to serious or fatal workplace injuries?

The data that helps to answer this data comes from the Bureau of Labor Statistics (BLS, ref. 3) and the United States Department of Labor-Occupational Safety and Healthy administration (OSHA) (ref. 1,2). The two data sources provided data on workplace injuries- each with data on 2019, which is the year I have chosen to focus on. Each has data on specific industry. After subsetting the OSHA data to this year, I was able to merge the data based on a NAICS (North American Industry Classification) code (ref.2), which classifies each industry under a code (or summary of injuries, based on the the BLS).

Each row of the data set contains a specific instance of a severe injury report. This comes from the OSHA data set. By merging the two data sets together, the BLS dataset added to each injury instance the number of fatal injuries overall based on the NAICS code- so each row also contains info about overall fatal injuries for their Industry classification, which is helpful to compare data info.

Each column of the data includes info on the date the specific severe injury occurred, the employer, the NAICS code, the corresponding industry name, and ID code (correspond the injury to filed paperwork), if a hospitalization or Amputation was necessary, the nature of the injury (amputation, fractures, crushing, etc.), the part of the body that was injured, an event title which explains how the injury occured, the source (and a secondary source if neccessary) of the injury, as well as the industry name and total fatal injuries (a death at the result of an injury, ref.3) in the year 2019.

Because both data sets comes from a census on workplace injuries, the data is meant to represent an entire population, not just a sample of a population, as a census is a study of all individual members in a population.

For the OSHA data set, which started in 2015, requires all employers to self report any severe injury to them (ref.1). This means their data collection method was via self-reporting from the population. The data set from the Bureau of Labor Statistics was done via a census, which is also done by self-reporting.

Furthermore, the data is appropriate to address the research question as it contains data on the population of interest, US workplace injuries, and helps to distinguish injuries based on the industry in which they occurred, which will allow me to conduct different graphical and numerical summaries that will allow me to compare different industries based on numbers of injuries and severity of the injuries.

Data Issues

Because the data is self-reported from both sources, this makes it possible that not every workplace injury is reported. This could lead to certain discrepancies in the data, especially if one or more industries is less likely to report of an industry. This issue also questions the status of the data as a population, though based off it being a census, it still makes the most sense to treat it as a population.

Numerical Summary

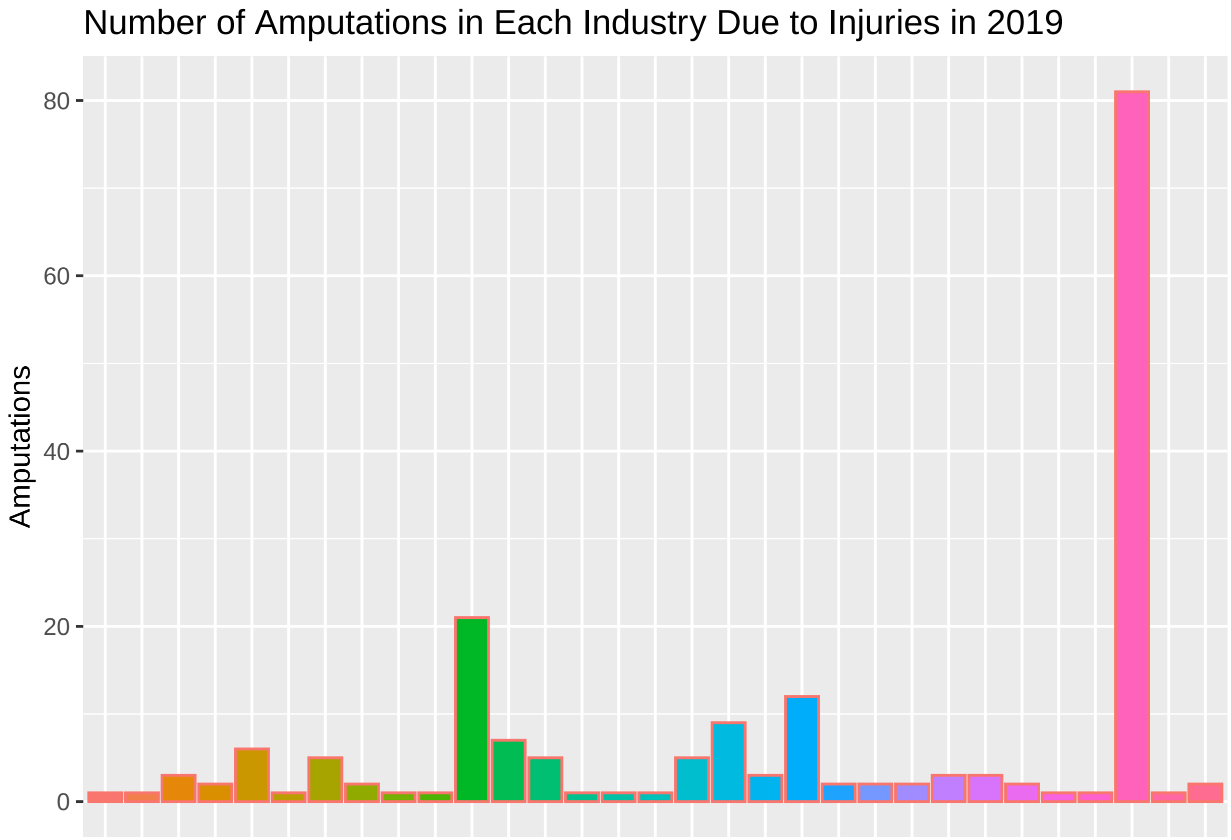
```
# Number of fatal injuries by industry
Industry_fatal<-distinct(Industry_fatal)
summary(as.numeric(Industry_fatal$Total.fatal.injuries..number.))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00    5.00   29.02   14.75  617.00
```

The above numerical summary is a 5 number summary (minimum, first quantile, median, mean, third quantile, and max) of the fatal injuries in each industry within the dataset. This is relevant as it gives us a base comparison of fatal injuries in the Industries we are looking at.

Graphical Summary

```
amp<- function(x){
  subset <- data_together[data_together$Industry.1. == x,]
  subset
  num<- sum(subset$Amputation, na.rm=T)
  num
}
vect<- unique(data_together$Industry.1.)
Amps<-sapply(vect, amp)
Amps<- as.numeric(Amps)
Industry<- vect
df1<- data.frame(Industry, Amps)
df1 <- df1[df1$Amps > 0,]
ggplot(df1, aes(x=Industry, y=Amps, color="black", fill=Industry))+ geom_bar(stat="ident
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



```
ggplot(df1, aes(x=Industry, y=Amps, color=Industry))+ geom_point(stat="identity") + them
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

Industry
● Armored car services
● Assisted living facilities for the elderly
● Automotive body, paint, and interior repair and maintenance
● Car washes
● Cattle feedlots
● Communication equipment repair and maintenance
● Construction, mining, and forestry machinery and equipment rental and leasing
● Continuing care retirement communities
● Custom computer programming services
● Deep sea freight transportation
● Drilling oil and gas wells
● Electric power distribution
● Employment placement agencies
● Farm labor contractors and crew leaders
● Finfish fishing
● Fossil fuel electric power generation
● Full−service restaurants
● General automotive repair
● General freight trucking, long−distance, less than tru
● General freight trucking, long−distance, truckload
● Limited−service restaurants
● Nonresidential property managers
● Nursery and tree production
● Oilseed and grain combination farming
● Residential property managers
● Security guards and patrol services
● Snack and nonalcoholic beverage bars
● Soil preparation, planting, and cultivating
● Support activities for oil and gas operations
● Truck transportation
● Utility system construction

The above graphical summary shows the number of amputations required in the year 2019 in all industries that required amputations. This is important as it shows the severity of the injuries by a qualitative factor that is easy to distinguish by- whether part of the body was need to be removed in order for the individual to survive. This gives us a way to start to answer the research question: Is there any industry or subset of industry that is most prone to serious or fatal workplace injuries?

Conclusions

One conclusion that can be drawn from the numerical summary is that fatal injury data is heavily impacted by outliers. The mean of the data is 29, whereas the median is 5. This is also supported by a max of 617, which means that there is an outlier in the dataset. Additionally, the five number summary gives us a good norm as to the normal amount of fatalities within each industry within the dataset. Due to the previously mentioned outlier, the median would be a better measure of center. This means that any industry with above 5 fatalities could be considered more dangerous than another.

Based on the bar graph on the number of amputees in each dataset, we can see that there is an extreme outlier in the Industry "Support activities for oil and gas operations". Additionally the other outlier in the data is "Drilling oil and gas wells". These are obviously the two professions within the data that have caused the most amputations, though without using a proportion of amputations to the amount of employees in a field, it is not possible to conclude that they are more "dangerous" than other industries.

References

1. https://www.osha.gov/severeinjury/
2. https://www.osha.gov/Establishment-Specific-Injury-and-Illness-Data
3. https://www.bls.gov/iif/oshcfoi1.htm