Analysis with R
Final project
Vajiheh Aghebati - May2020

Introduction

Drowsy Driving

"According to the National Highway Traffic Safety Administration, every year about 100,000 police-reported crashes involve drowsy driving. These crashes result in more than 1,550 fatalities and 71,000 injuries. The real number may be much higher, however, as it is difficult to determine whether a driver was drowsy at the time of a crash."
"NHTSA's census of fatal crashes and estimate of traffic-related crashes and injuries rely on police and hospital reports to determine the incidence of drowsy-driving crashes. NHTSA estimates that in 2017, 91,000 police-reported crashes involved drowsy drivers. These crashes led to an estimated 50,000 people injured and nearly 800 deaths. But there is broad agreement across the traffic safety, sleep science, and public health communities that this is an underestimate of the impact of drowsy driving."

This report presents data regarding sleep/fatigue related crashes as it currently exists in "Open Data Pennsylvania" database.
This data base provide Crash data reported to the Pennsylvania Department of Transportation. Includes data involving drivers, passengers, and motor vehicles for researching highway safety, From 1997 to 2017.

Asleep/Fatigue-related Crash is a crash in which the driver was reported sleepy based on the police accident report.

In this report I worked on years 2015 to 2017.

According to this data set, drowsy driving was involved in 2.1 percents of all crashes result in injury or death, from 2015 to 2017 in PA.

Question:

Are Asleep/Fatigue related crashes more likely between young drivers than older drivers?

Methodology

The data set includes a total of 382181 rows of data for crashes happened between 2015 to 2017 and 180 column of information for each record.
The columns that I work mostly on them looked like table below:

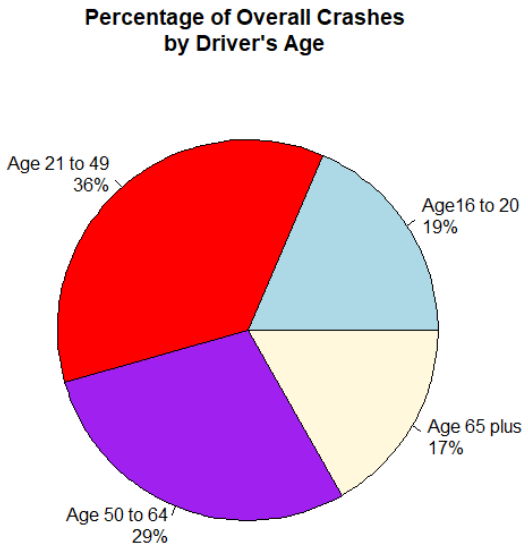
Driver age 16	**Driver age 17**	**Driver age 18**	**Driver age 19**	**Driver age 19**	**Driver age 20**	**Driver age 50to64**	**Driver age 65to74**	**Driver age 75+**	**Fatigue /Asleep Driver**	**Injury or Fatal**	**Hour of the day**
Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No	numeric

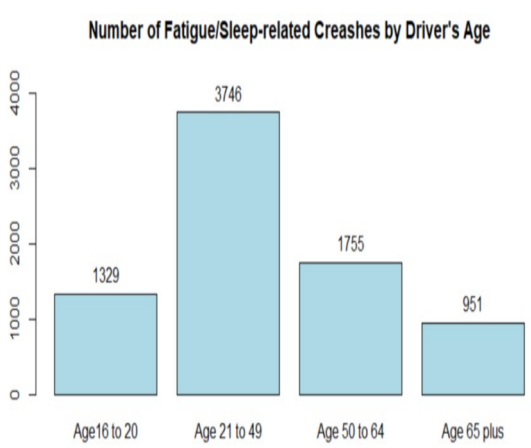
So I decided to categorize the drivers by age groups as described below:

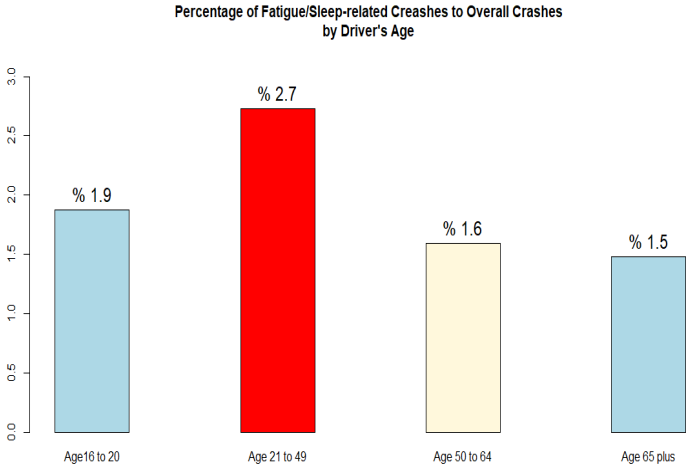
Age16to20 (Young)	Age 21to49 *	Age50to64	Age65+ (Old)

***** Age 21to49 was nor specified in the data set, so I assumed that every record which not included in any of other groups, is related to a driver of age 21 to 49.

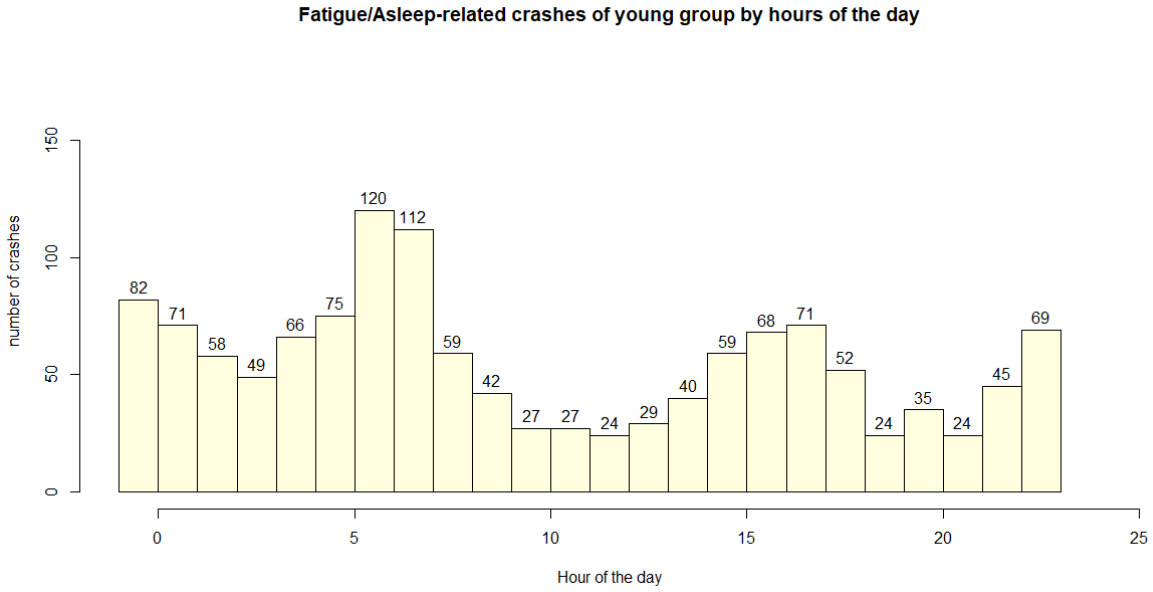
In order to answer the “Question” I ran a Hypothesis Test (Proportion Test) on the two group of so-called young and old drivers to compare the likelihood of involving in a fatigue/asleep related crash.

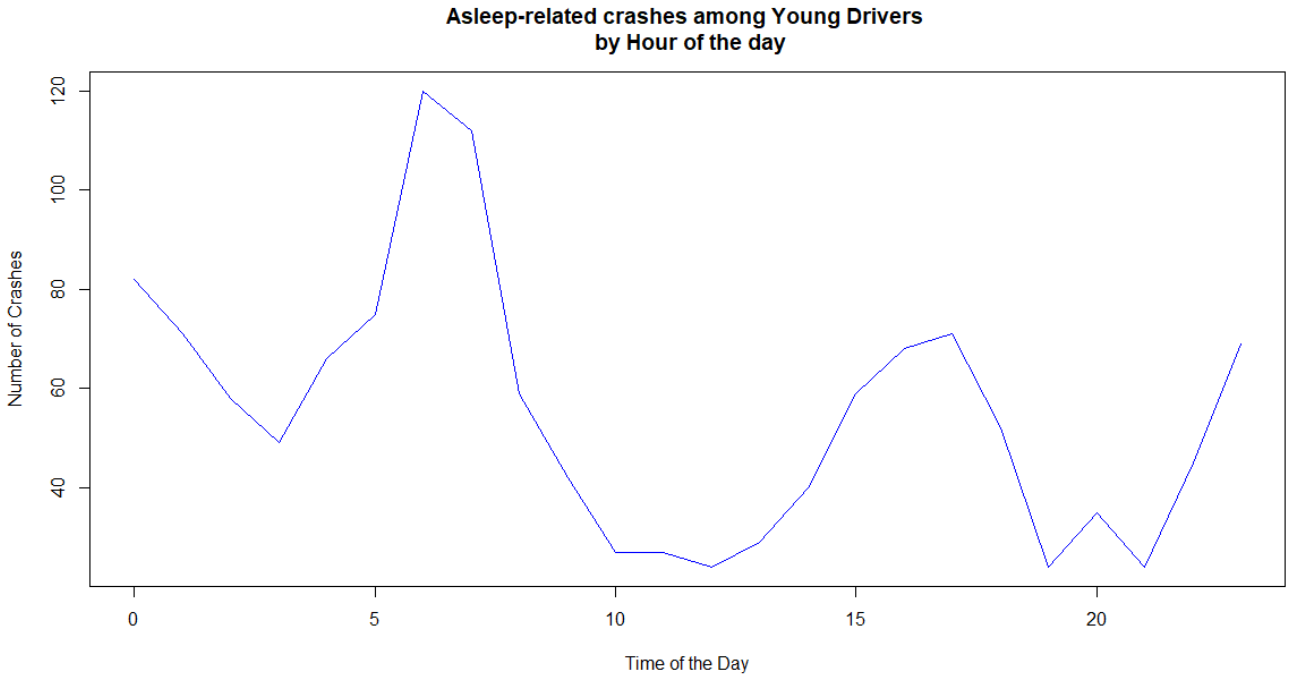






The following plot and table shows the Fatigue/asleep related crashes between young drivers during the hours of the day, and key statistics of it.





> summary()

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 24.00   33.50   55.00   55.33   69.50  120.00
```

Mean =55.3, shows that in each hour of the day about 55 crashes happened in average.

According to the calculated statistics, here some question about the data which answered in R:

-In which hour of the day maximum crashes happened?(mode)
6 am
-In which hour of the day minimum crashes happened?
12,19 and 21 pm
-In which hours of the day crash numbers were greater than the average Crash number?
0,1,2,4,5,6,7,8,15,1617,23
-Does the number of crashes correlated to the time of the day?
Results shows cor = -0.45 so in means negatively weak to moderate correlated

Hypothesis Testing:

In order to answer this analysis question I ran a proportion test in R as follows:

Null Hypothesis: The proportion of Young drivers in sleep-relates crashes are “equal to” the proportion of Old drivers.
Alternative Hypothesis: The proportion of Young drivers in sleep-relates crashes are “ not equal to” the proportion of Old drivers.

> table :

```
              sleeprelated    nonsleeprelated
young          1329            69493
old             951            63195
```

> prop.test(table,correct = FALSE) :

```
    2-sample test for equality of
        proportions without continuity
        correction

X-squared = 31.459, df = 1, p-value =
2.037e-08
alternative hypothesis: two.sided
95 percent confidence interval:
 0.002571066 0.005308536
sample estimates:
    prop 1     prop 2
0.01876536 0.01482555
```

Interpretation of the results

the p-value is very small so we can reject the Null Hypothesis (the proportions are equal),indicates that the proportions of the characteristic studied are statistically significantly different in the 2 groups.

Also since the proportion in young is greater than the other one so we can conclude the young drivers **more likely** to have a sleep-related crash than old Drivers.
If we run a one-sided proportion test with “alternative=”greater” ” the p-value is still very small which confirm our conclusion.

```
X-squared = 31.459, df = 1, p-value =
1.018e-08
alternative hypothesis: greater
95 percent confidence interval:
 0.002791123 1.000000000
sample estimates:
    prop 1     prop 2
0.01876536 0.01482555
```

conclusion

According to this analysis and the data set for vehicle crashes between 2015 to 1017 in PA, we can conclude drivers aged 16 to 20 are more likely to involve in a Fatigue/Asleep related crash than the drivers ages 65+ .

References

1- h(https://data.pa.gov/Public-Safety/Crash-Incident-Details-CY-1997-Current-Annual-Coun/dc5b-gebx)

2- https://www.nhtsa.gov/risky-driving/drowsy-driving

3- https://www.nsc.org/road-safety/safety-topics/fatigued-driving