Assignment-6

Saurabh Yelne

4/6/2018

Question 1

Chart Name	Tool	Dataset Name
Radar chart	R	mtcars dataset present in R
Back to back bar chart	R	Police killings data set given during midterm, sheet 1
Treemap	R	City of london data set from the gov.uk website
Dendrogram	R	Flare dataset https://cdn.rawgit.com/christophergandrud/networkD3/,"master
Parallel coordinates	R	mtcars dataset present in R
Correlations heat map	R	wine quality dataset from UCI repository
Chord diagram	R	flights dataset from nycflights13 library
Shanky diagram	R	Police killings data set given during midterm, sheet 1
Bump chart	R	UniversityData from kaggle
Dot map	R	Crime dataset from ggmap library

Table

Question 2

Answered within every chart

1. Radar chart

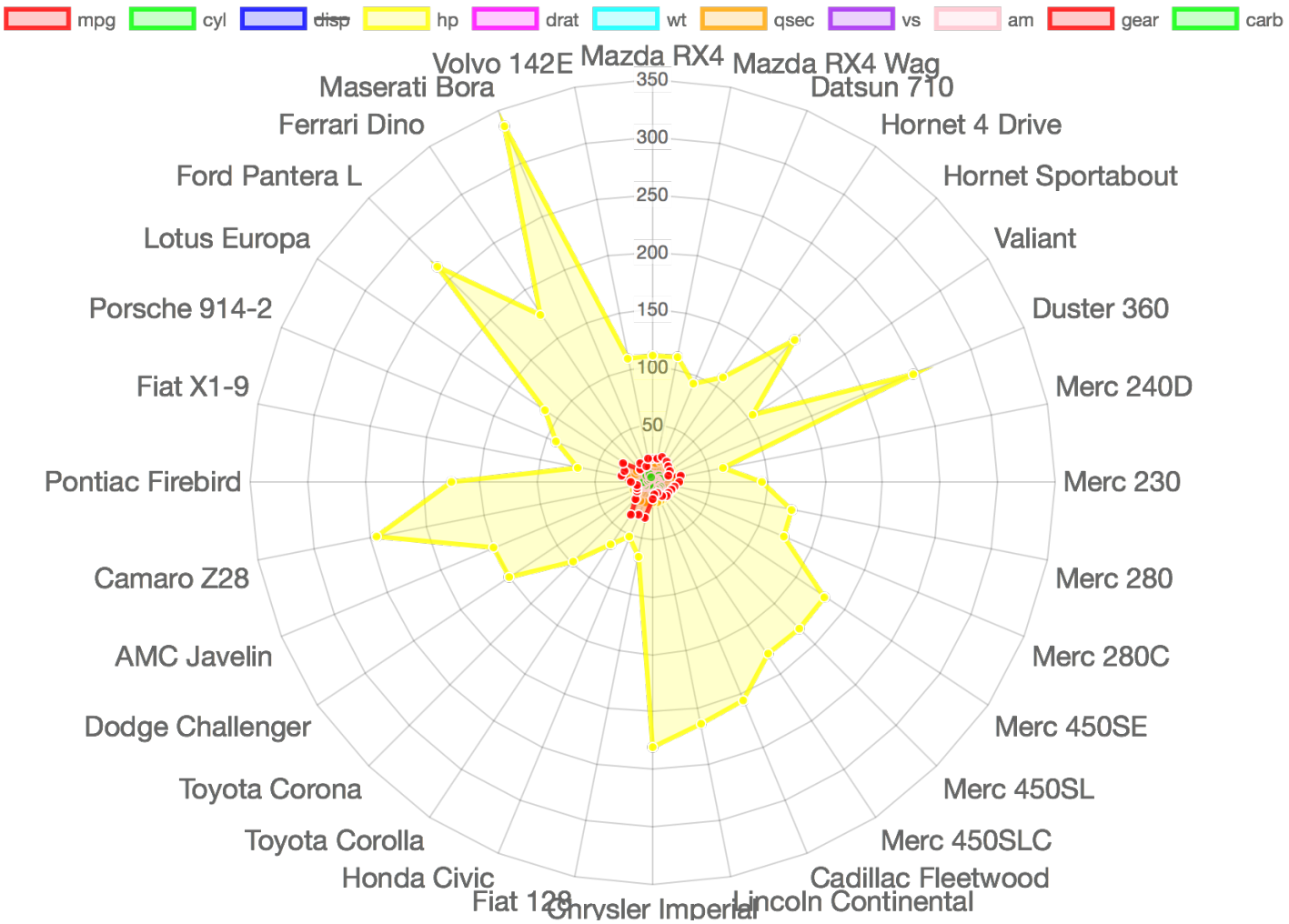
a. Briefly describe about the dataset This dataset is about different car models and their characteristics like mpg, # no cylinders etc

b. Precisely explain why the chart is appropriate for the dataset Radar Charts comparies various attributes amongst different variables. Thus we can compare different cars and their attributes with this chart.

c. Write two or three interesting observations one can make from the chart

1. This chart is interactive
2. Most of the models are 8 cylinders
3. Min number of gear is 3

```
cars<-tibble::rownames_to_column(mtcars,var='Cars')
chartJSRadar(cars)
```



2. Back to back chart

a. Briefly describe about the dataset This dataset gives the people killed by police. Gender and race information is given

b. Precisely explain why the chart is appropriate for the dataset To compare gender and their race back to back chart is very helpful in getting insights

c. Write two or three interesting observations one can make from the chart.

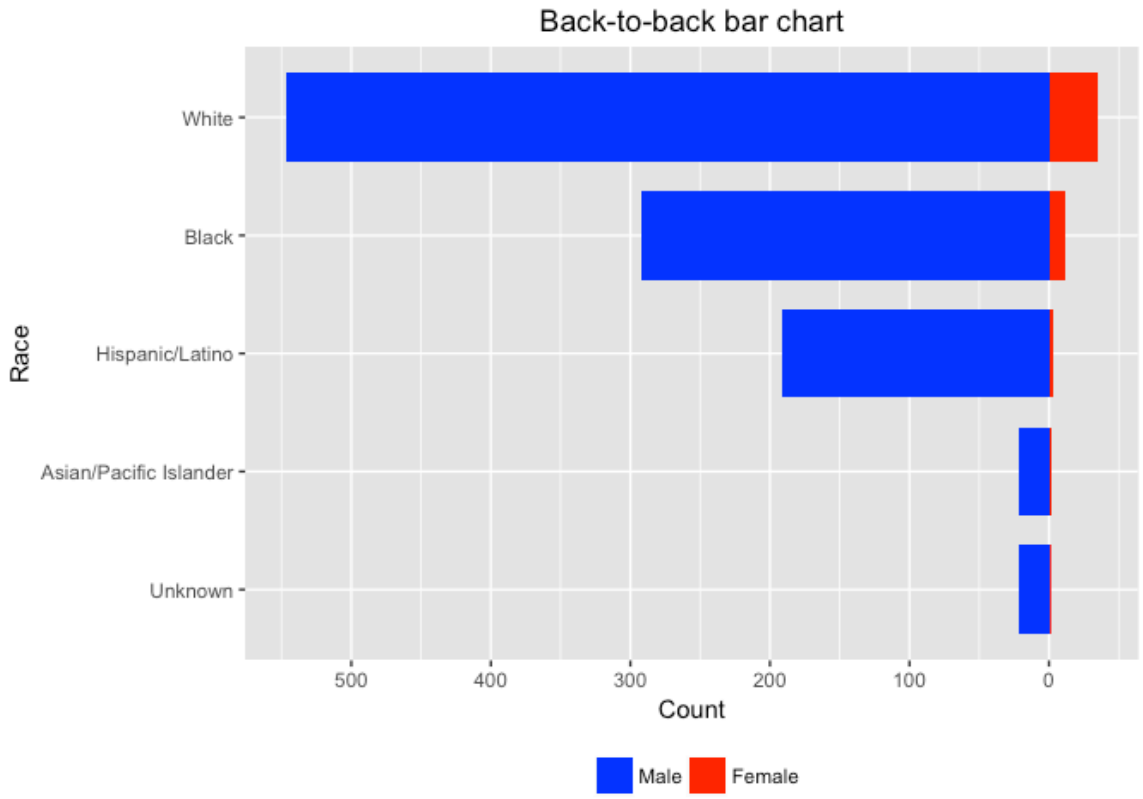
1. Men are mostly killed across all the races compared
2. Most women killed are white

```
killings<-readxl::read_xlsx('PoliceKillings.xlsx',sheet = 1)
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018c.1.(
```

```
bar<-killings%>%group_by(raceethnicity,gender)%>%summarise(freq=n())%>%mutate(
female <-
  bar %>%
  filter(gender == "Female") %>%
  arrange(freq)
order<-female$raceethnicity
order<-as.factor(order)
p <-
  bar %>%
  ggplot(aes(x = raceethnicity, y = freq, group = gender, fill = gender)) +
  geom_bar(stat = "identity", width = 0.75) +
  coord_flip() +scale_x_discrete(limits = order)
  p1<-p+scale_y_continuous(breaks = seq(-600, 150, 100),
                     labels = abs(seq(-600, 150, 100)))+
    labs(x = "Race", y = "Count", title = "Back-to-back bar chart") +
  theme(legend.position = "bottom",
        legend.title = element_blank(),
        plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill =  "grey90")) +
  # reverse the order of items in legend
   #guides(fill = guide_legend(reverse = TRUE)) +
  # change the default colors of bars
  scale_fill_manual(values=c("red", "blue"),
                    name="",
                    breaks=c("Male", "Female"),
                    labels=c("Male", "Female"))
```

```
print(p1)
```



3. Treemap

a. Briefly describe about the dataset

The dataset tell about the nature of crimes in the city of London

b. Precisely explain why the chart is appropriate for the dataset

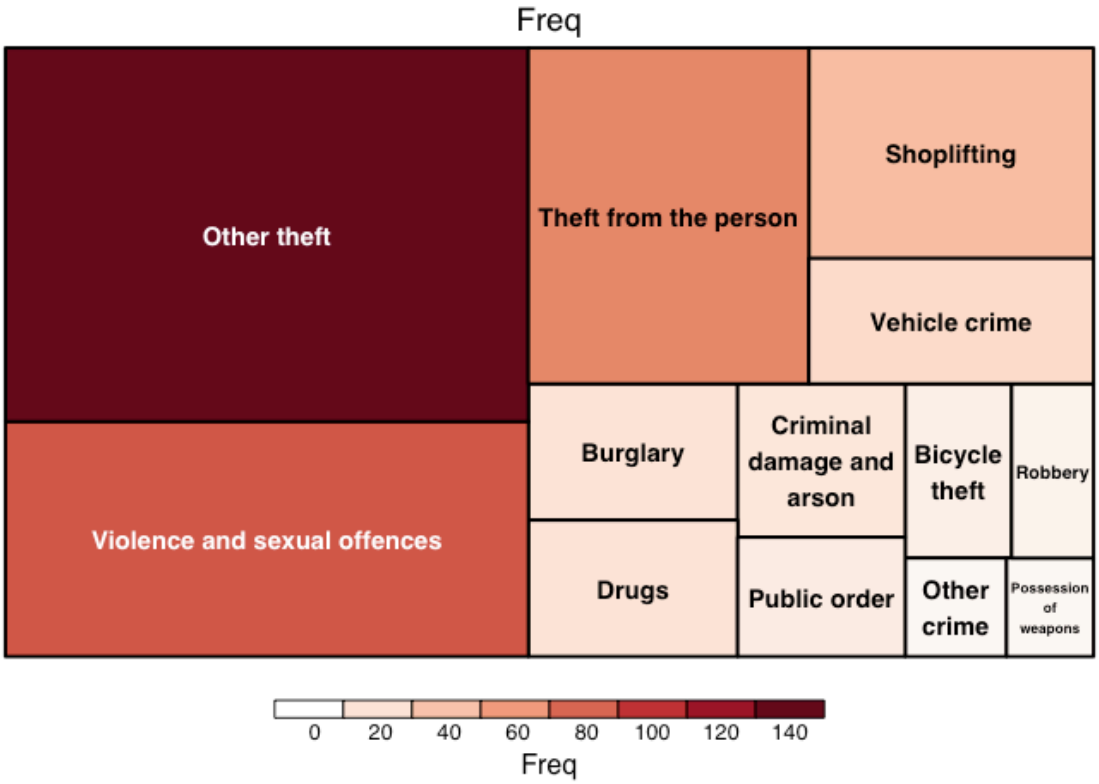
To see what nature of crime has what proportion, tree chart is of great help

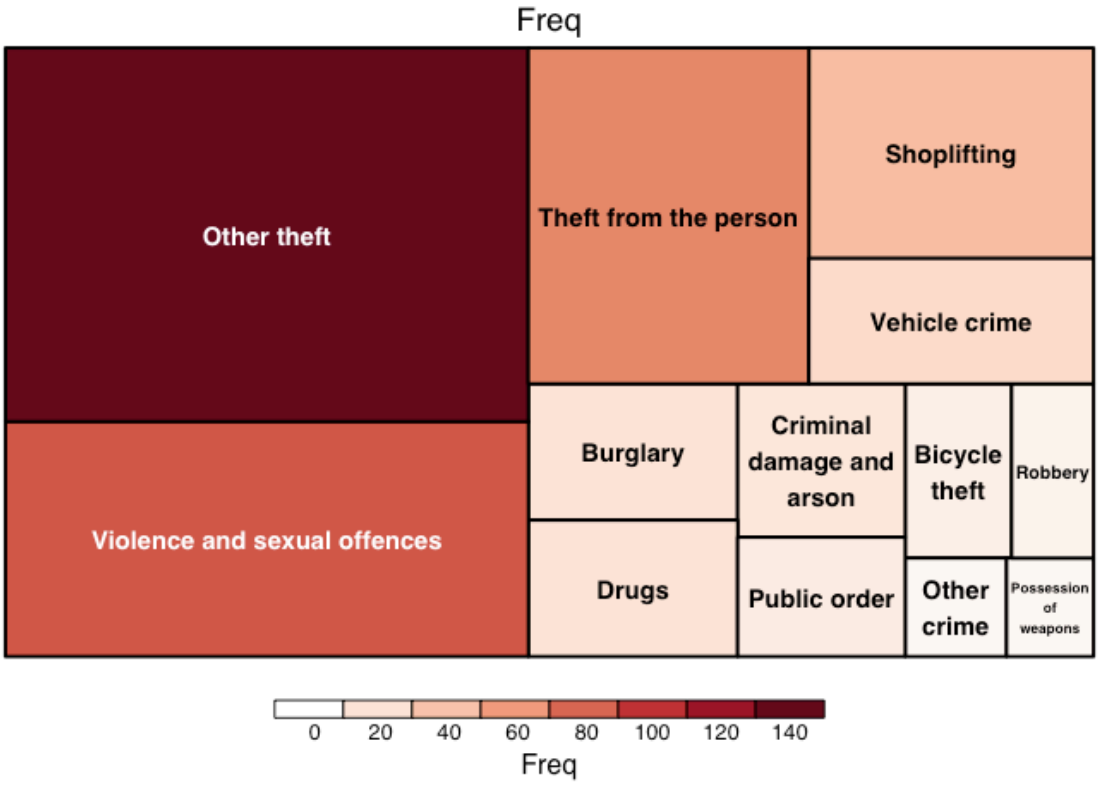
c. Write two or three interesting observations one can make from the chart.

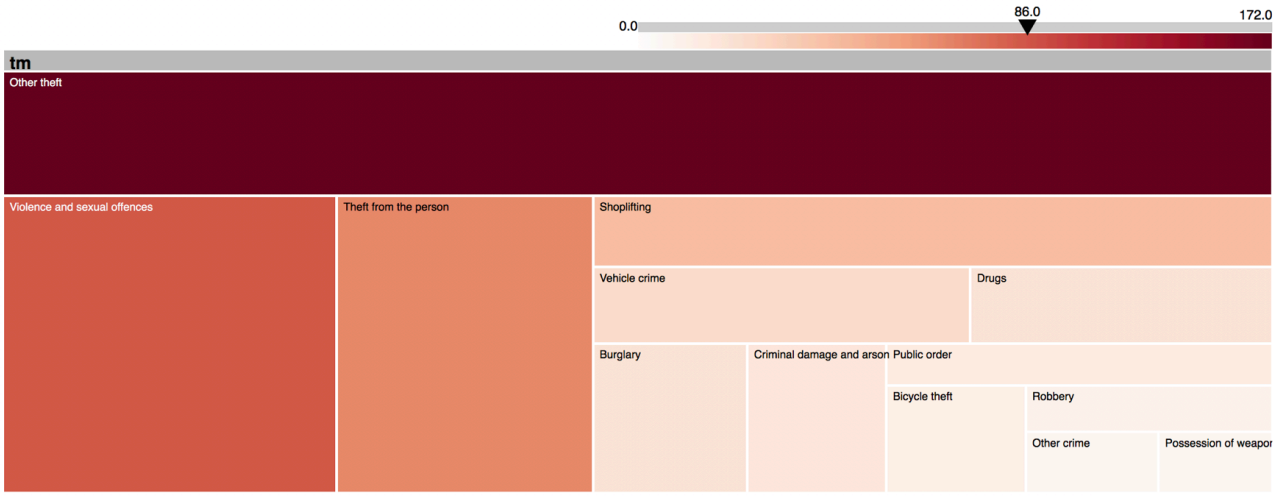
1. Most crimes are related to thefts
2. Possession of weapons is the least amount of crime done

```
london<-read.csv('city-of-london-street.csv')
london_map<-london%>%select(Crime.type,Longitude,Latitude,Location)
tree<-data.frame(table(london_map$Crime.type))
tm<-treemap(
```

```
tree,
index=("Var1"),
vSize="Freq",
vColor = 'Freq',
type = 'value',
palette="-RdGy",
range=c(5,150)
)
d3tree3(tm) #gives an html interactive widget output
```




Freq
Other theft
Theft from the person
Shoplifting
Vehicle crime
Violence and sexual offences
Burglary
Criminal damage and arson
Bicycle theft
Robbery
Drugs
Public order
Other crime
Possession of weapons
0
20
40
60
80
100
120
140
Freq



4. Dendogram

a. Briefly describe about the dataset

Flare dataset is a list of cluster.

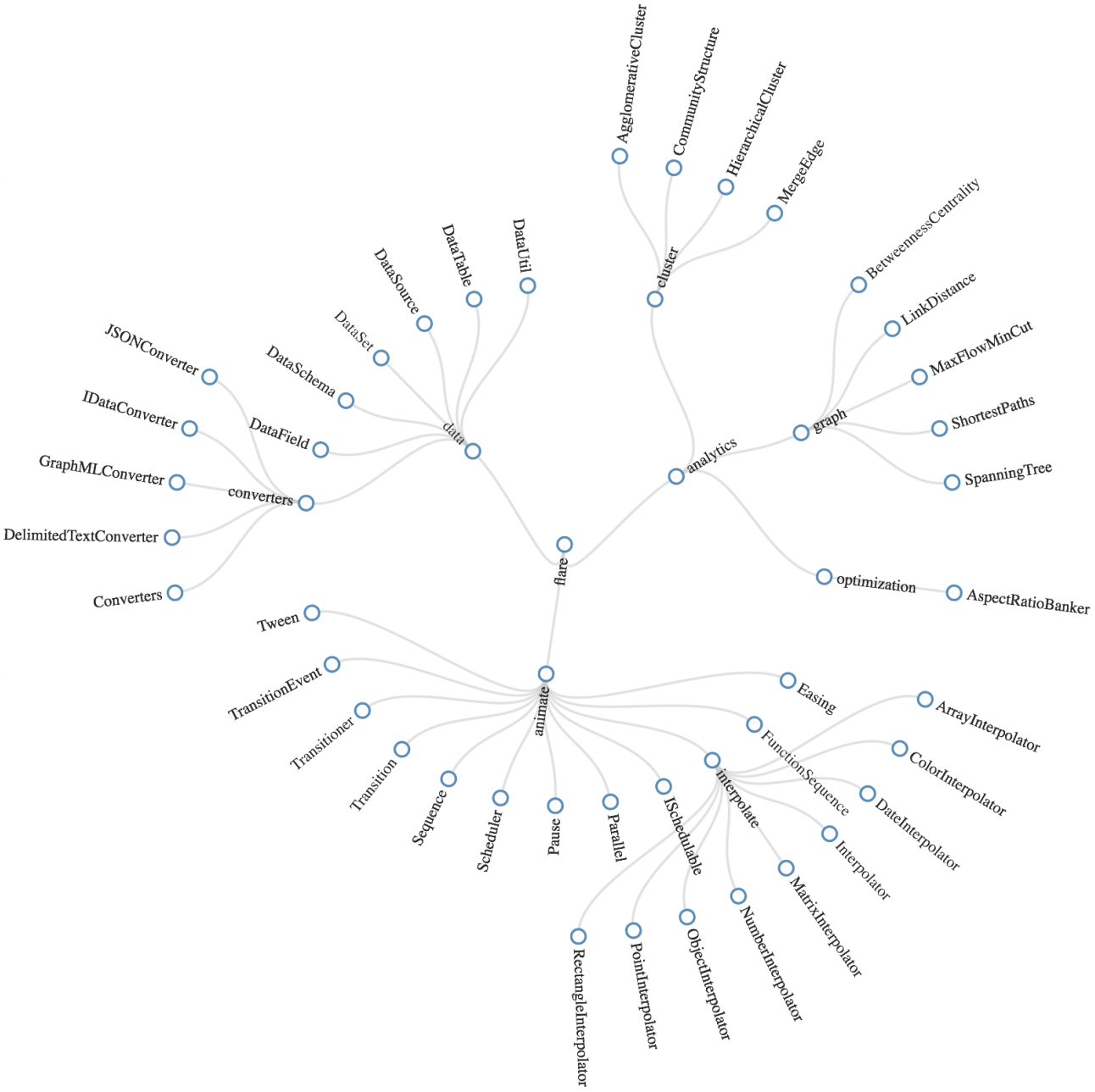
b. Precisely explain why the chart is appropriate for the dataset

To see how clusters are formed, dendograms are the best charts.

c. Write two or three interesting observations one can make from the chart.

1. There are three main clustures namely analytics, animate and data
2. The main clusters are then further divided into sub clusters
3. This chart is interactive

```
URL <- paste0(
"https://cdn.rawgit.com/christophergandrud/networkD3/",
"master/JSONdata//flare.json")
Flare <- jsonlite::fromJSON(URL, simplifyDataFrame = FALSE)
Flare$children = Flare$children[1:3]
radialNetwork(List = Flare, fontSize = 10, opacity = 0.9)
```



5.Parallel co-ordinates

a. Briefly describe about the dataset

This dataset is about different car models and their characteristics like mpg, # no cylinders etc

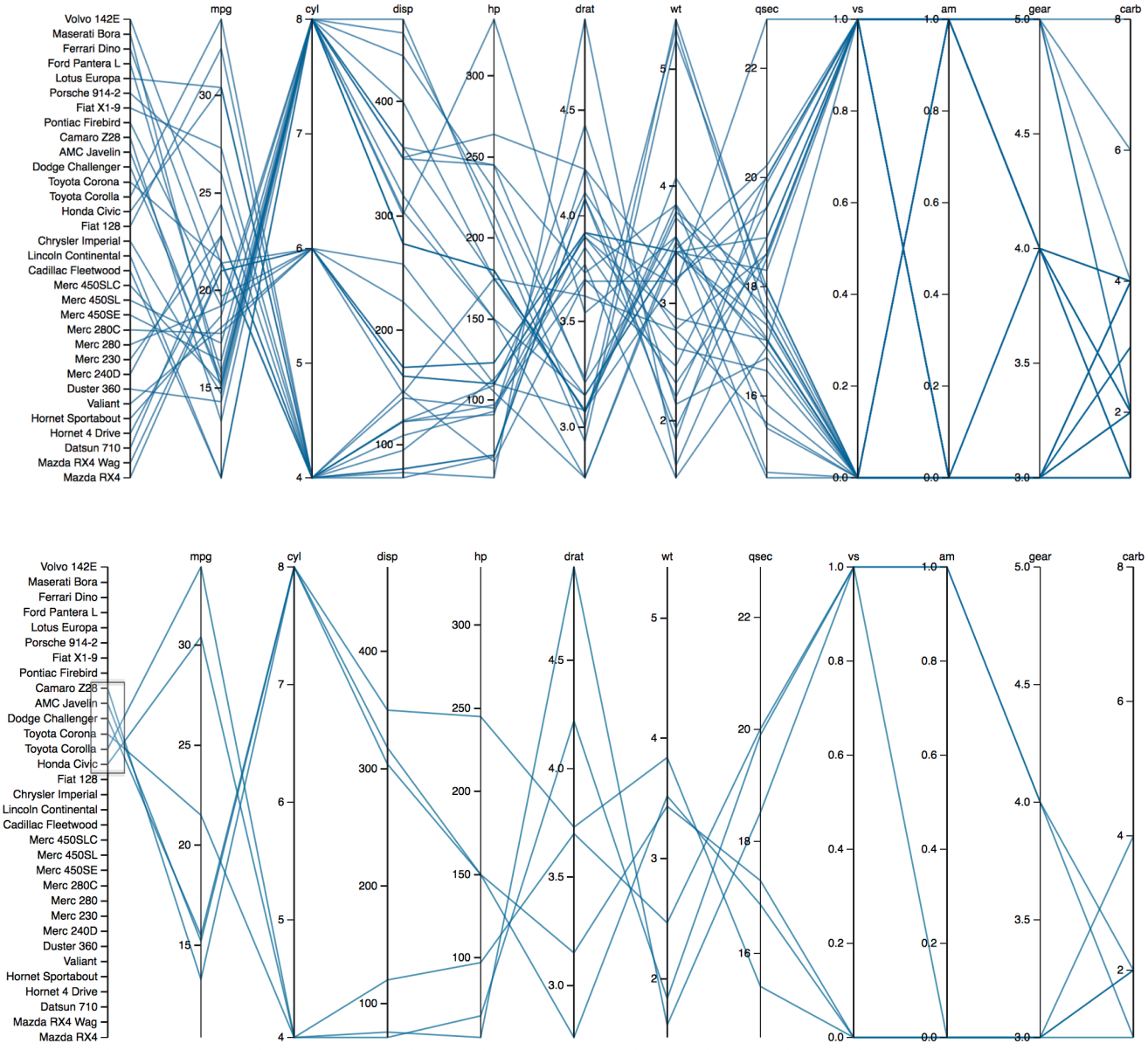
b. Precisely explain why the chart is appropriate for the dataset

To compare various models with their attributes, parallel co-ordinate is a good chart

c. Write two or three interesting observations one can make from the chart.

1. This chart is interactive and we can use filters for every attribute axis

```
parcoords(mtcars
, brushMode  = "1d-axes"
, reorderable = TRUE)
```



Tree-Interactive

6.Correlation: Heatmap

a. Briefly describe about the dataset This dataset gives information about the white wine

b. Precisely explain why the chart is appropriate for the dataset We want to find correlation between various attributes of the white wine and thus this chart is useful

c. Write two or three interesting observations one can make from the chart.

1. Red color means strong positive correlation
2. Blue color is strong negative correlation

```
#installing the data file
wine<-read.delim("winequality-white.csv",sep = ";")

#Check for na's
any(is.na(wine))  #no na's in the dataset
```

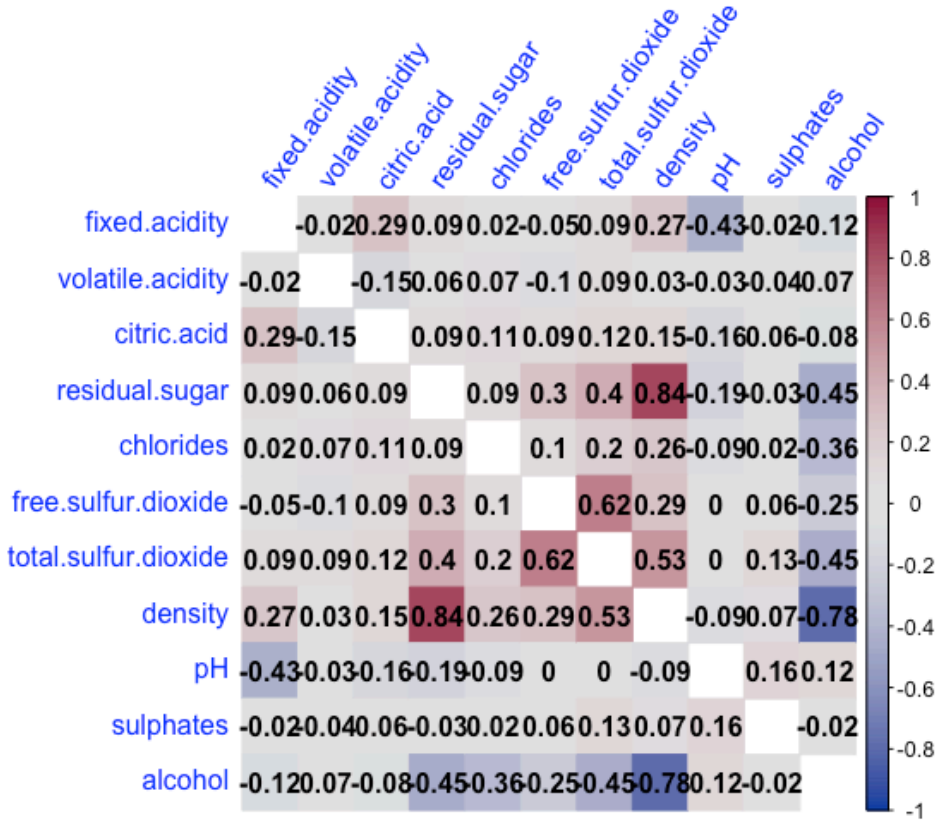
```
## [1] FALSE
```

```
#removing quality column
matrix<-wine[-12]

#correlation matrix
m<-cor(matrix)

#choosing a color palette
pal<-choose_palette()
```

```
## Warning: running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framew
```

```
## Loading required namespace: dichromat
```

```
#visualizing the correlation matrix
corrplot(m,method = "color",col=pal(200),
         addCoef.col = "black",diag = F,tl.col = "blue",tl.srt = 55,addshade =
```



7.Chord Diagram

a. Briefly describe about the dataset

Flight datasets gives information about the various carriers, their delays, origin and destination airports etc for the year 2013

b. Precisely explain why the chart is appropriate for the dataset To see air traffic is moving, which is the origin airport and the destination airport, this chart is very useful. I am seeing the first 100 entries

c. Write two or three interesting observations one can make from the chart.

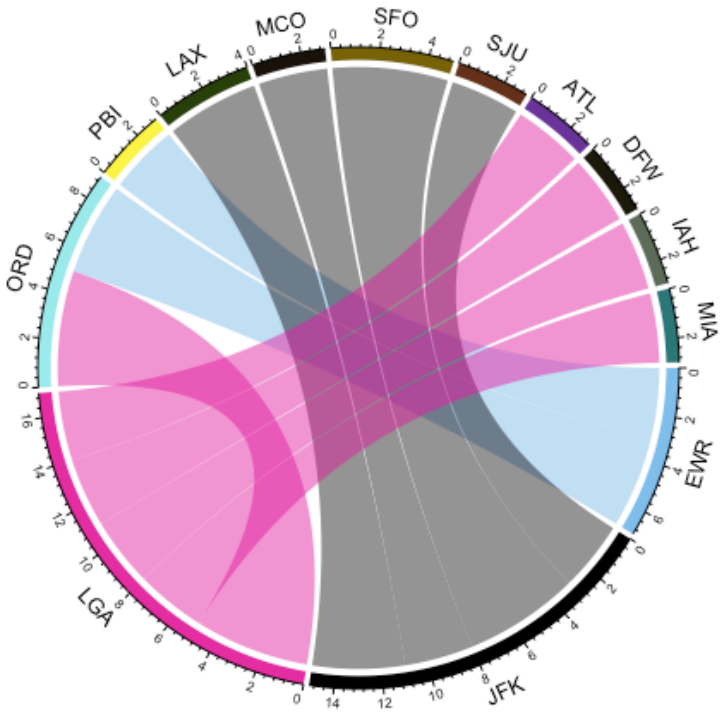
1. LGA and JFK are two major airports from where most flight are flying out
2. ORD airport is recieving the most number of flights

```
library(nycflights13)
data("flights")
flights
```

```
## # A tibble: 336,776 x 19
```

```
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_t
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <i
##  1  2013     1     1      517            515        2.      830
##  2  2013     1     1      533            529        4.      850
##  3  2013     1     1      542            540        2.      923
##  4  2013     1     1      544            545       -1.     1004
##  5  2013     1     1      554            600       -6.      812
##  6  2013     1     1      554            558       -4.      740
##  7  2013     1     1      555            600       -5.      913
##  8  2013     1     1      557            600       -3.      709
##  9  2013     1     1      557            600       -3.      838
## 10  2013     1     1      558            600       -2.      753
## # ... with 336,766 more rows
```

```
travel<-flights%>%select(origin,dest)%>%head(100)
travel1<-plyr::count(travel, c("origin", "dest"))
travel1<-travel1%>%filter(freq>2)
circlize::chordDiagram(travel1)
```



8.Sankey

a. Briefly describe about the dataset This dataset gives the people killed by police. Gender and race information is given

b. Precisely explain why the chart is appropriate for the dataset In order to understand how the males and females are killed by the police, sankey diagram is useful chart

c. Write two or three interesting observations one can make from the chart.

1. Male and females are mostly killed by the gunshot
2. A very small amount of females are struck by truck and killed

```
killings<-readxl::read_xlsx('PoliceKillings.xlsx',sheet = 1)
head(killings)
```

```
## # A tibble: 6 x 14
##       P name              age   gender raceethnicity          month     day
##   <dbl> <chr>             <chr> <chr>  <chr>                  <chr>   <dbl>
## 1    2. Matthew Ajibade   22    Male   Black                  January    1.
## 2    4. Lewis Lembke      47    Male   White                  January    2.
## 3    7. Tim Elliott       53    Male   Asian/Pacific Islander January    2.
## 4    5. Michael Kocher Jr 19    Male   White                  January    3.
## 5    6. John Quintero     23    Male   Hispanic/Latino        January    3.
## 6    8. Matthew Hoffman   32    Male   White                  January    4.
```

```
male<-killings%>%filter(gender=='Male')
female<-killings%>%filter(gender=='Female')
table(male$classification)
```

```
##
##  Death in custody          Gunshot            Other Struck by vehicle
##                40              974                  1                27
```

```
table(female$classification)
```

```
##
##  Death in custody          Gunshot Struck by vehicle            Taser
##                 1               44                  7                1
```

```
kills <- data.frame(gender=c(
  rep('Male',4),rep('Female',4)),
```

```
  death_classification=c(
    "Death in custody ","Gunshot ","Struck by vehicle  ", "Taser "),
  weights=c(
    c( 40,974,27,49),
    c(  1,44,7,1)))


plot(
  gvisSankey(kills, from="gender",
                  to="death_classification", weight="weights",
                  options=list(
                    height=250,
                    sankey="{link:{color:{fill:'lightblue'}}}"
                    ))
  )
```

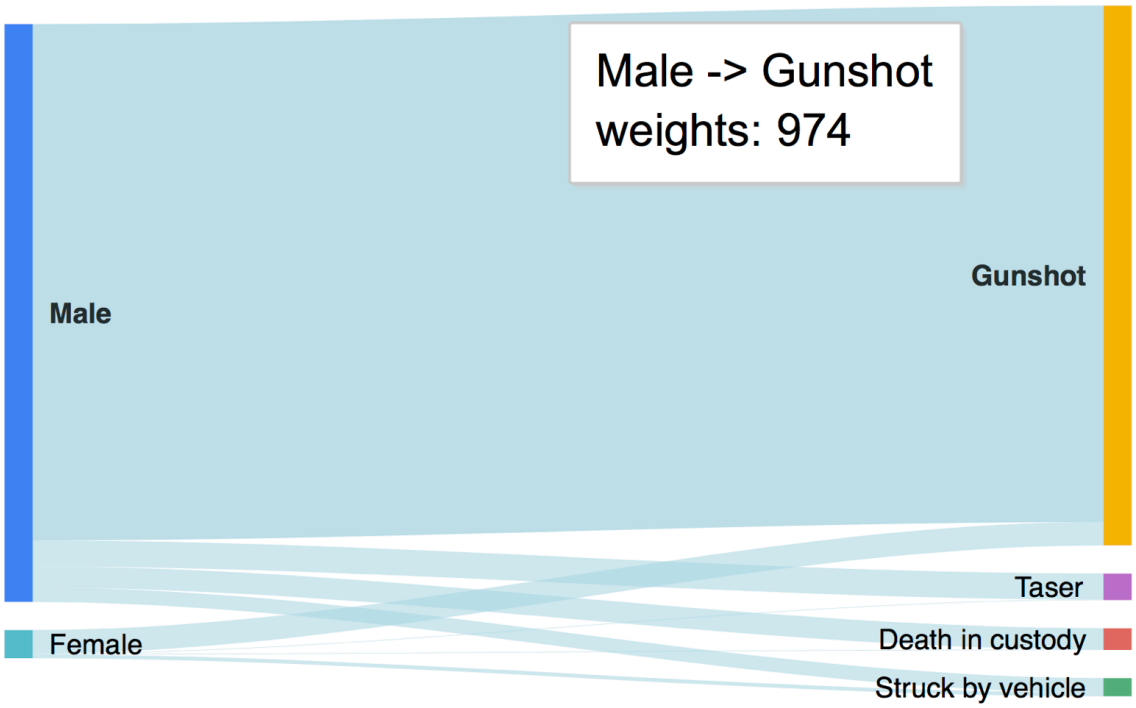
```
## <!-- Sankey generated in R 3.4.2 by googleVis 0.6.2 package -->
## <!-- Sat Apr  7 20:40:38 2018 -->
##
##
## <!-- jsHeader -->
## <script type="text/javascript">
##
## // jsData
## function gvisDataSankeyID32756112efae () {
## var data = new google.visualization.DataTable();
## var datajson =
## [
##  [
## "Male",
## "Death in custody ",
## 40
## ],
## [
## "Male",
## "Gunshot ",
## 974
## ],
## [
## "Male",
## "Struck by vehicle  ",
## 27
## ],
## [
## "Male",
```

```
## "Taser ",
## 49
## ],
## [
## "Female",
## "Death in custody ",
## 1
## ],
## [
## "Female",
## "Gunshot ",
## 44
## ],
## [
## "Female",
## "Struck by vehicle  ",
## 7
## ],
## [
## "Female",
## "Taser ",
## 1
## ]
## ];
## data.addColumn('string','gender');
## data.addColumn('string','death_classification');
## data.addColumn('number','weights');
## data.addRows(datajson);
## return(data);
## }
##
## // jsDrawChart
## function drawChartSankeyID32756112efae() {
## var data = gvisDataSankeyID32756112efae();
## var options = {};
## options["width"] = 400;
## options["height"] = 250;
## options["sankey"] = {link:{color:{fill:'lightblue'}}};
##
##
##     var chart = new google.visualization.Sankey(
##     document.getElementById('SankeyID32756112efae')
##     );
##     chart.draw(data,options);
##
##
```

```
## }
##
##
## // jsDisplayChart
## (function() {
## var pkgs = window.__gvisPackages = window.__gvisPackages || [];
## var callbacks = window.__gvisCallbacks = window.__gvisCallbacks || [];
## var chartid = "sankey";
##
## // Manually see if chartid is in pkgs (not all browsers support Array.index
## var i, newPackage = true;
## for (i = 0; newPackage && i < pkgs.length; i++) {
## if (pkgs[i] === chartid)
## newPackage = false;
## }
## if (newPackage)
##   pkgs.push(chartid);
##
## // Add the drawChart function to the global list of callbacks
## callbacks.push(drawChartSankeyID32756112efae);
## })();
## function displayChartSankeyID32756112efae() {
##   var pkgs = window.__gvisPackages = window.__gvisPackages || [];
##   var callbacks = window.__gvisCallbacks = window.__gvisCallbacks || [];
##   window.clearTimeout(window.__gvisLoad);
##   // The timeout is set to 100 because otherwise the container div we are
##   // targeting might not be part of the document yet
##   window.__gvisLoad = setTimeout(function() {
##   var pkgCount = pkgs.length;
##   google.load("visualization", "1", { packages:pkgs, callback: function() {
##   if (pkgCount != pkgs.length) {
##   // Race condition where another setTimeout call snuck in after us; if
##   // that call added a package, we must not shift its callback
##   return;
## }
## while (callbacks.length > 0)
## callbacks.shift()();
## } });
## }, 100);
## }
##
## // jsFooter
## </script>
##
## <!-- jsChart -->
## <script type="text/javascript" src="https://www.google.com/jsapi?callback=d
```

```
##
## <!-- divChart -->
##
## <div id="SankeyID32756112efae"
##   style="width: 400; height: 250;">
## </div>
```

Sankey-Interactive Diagram Link



Data: kills • Chart ID: SankeyID10d4e18e2b310 • googleVis-0.6.2
R version 3.4.2 (2017-09-28) • Google Terms of Use • Documentation and Data Policy

Sankey Diagram Screenshot

9.Bump Chart

a. Briefly describe about the dataset The university dat gives world rankings of universities from 2012 to 2015

b. Precisely explain why the chart is appropriate for the dataset To see how university rankings are changing over time, bump chart is useful. I compared top 5 universities

c. Write two or three interesting observations one can make from the chart.

1. Harvard is maintaining its ranking on 1st position for all the years

2. Oxford moved from 3rd rank to 5th rank in year 2014.

```
uni<-read_csv('UniversityData.csv')
```

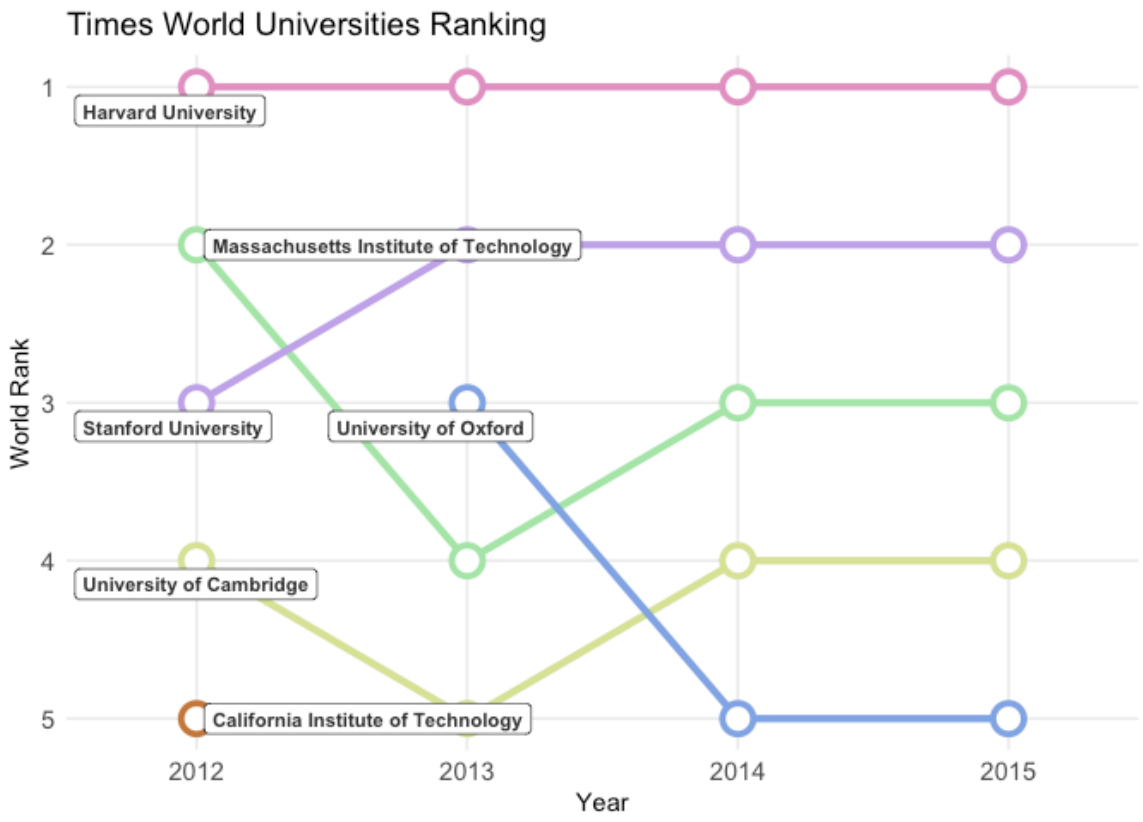
```
## Parsed with column specification:
## cols(
##   world_rank = col_integer(),
##   institution = col_character(),
##   country = col_character(),
##   national_rank = col_integer(),
##   quality_of_education = col_integer(),
##   alumni_employment = col_integer(),
##   quality_of_faculty = col_integer(),
##   publications = col_integer(),
##   influence = col_integer(),
##   citations = col_integer(),
##   broad_impact = col_integer(),
##   patents = col_integer(),
##   score = col_double(),
##   year = col_integer()
## )
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
```

```
## Warning: 4 parsing failures.
## row # A tibble: 4 x 5 col     row col         expected           actual fil
```

```
rank<-uni%>%select(world_rank,institution,year)%>%filter(world_rank<=5)
timesfirstAppearance <- rank[!duplicated(rank$institution) & rank$year != 2015
pal1 = c("#c57c3c", "#e392c2", "#a5e7a8", "#bea3ea", "#d7e298", "#81a4e3", "#a
         "#d6a16d", "#62d9f3", "#eb9189", "#3ec1c8", "#e1a6b6", "#7fe3c5", "#e

g1 <- ggplot(data=rank, aes(x=year, y=world_rank)) +
geom_line(aes(colour=institution), size=1.5) +
geom_point(shape = 21, stroke = 2, size=5, fill = "white", aes(colour=institut
geom_label_repel(data = timesfirstAppearance, aes(label=institution), size=3,
scale_y_reverse(lim=c(5,1), breaks = scales::pretty_breaks(n = 5)) +
scale_x_continuous(expand = c(.12, .12), breaks = scales::pretty_breaks(n = 4)
ggtitle('Times World Universities Ranking') +
xlab("Year") +
ylab("World Rank") +
theme_minimal() +
```

```
theme_bw() +
scale_colour_manual(values=pal1) +
theme(panel.background = element_rect(fill = '#ffffff'),
plot.title = element_text(size=14), legend.title=element_blank(),
axis.text = element_text(size=11), axis.title=element_text(size=11),
panel.border = element_blank(), legend.position='none',
panel.grid.minor.x = element_blank(), panel.grid.minor.y = element_blank(),
axis.ticks.x=element_blank(), axis.ticks.y=element_blank())

print(g1)
```



10.Dot Map

a. Briefly describe about the dataset Crime dataset gives information of various crimes in the city of Houston

b. Precisely explain why the chart is appropriate for the dataset To see what is the nature of crime in a particular part of the city and try to understand its density, dot map is used

c. Write two or three interesting observations one can make from the chart.

1. The density of aggrevated assault and robbery is the most across the area under observation 2.The concentration increases in the SW part near grey hound bus station

```
voilent_crimes<-crime%>%filter(offense!='auto theft'& offense!='theft'&offense
HoustonMap<-qmap('houston', zoom = 14)
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=houston
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?add
```

```
HoustonMap +geom_point(aes(x = lon, y = lat, colour = offense),data = voilent_
```