Week 9 & 10 Assignment - Python

Name: Kesav Adithya Venkidusamy
Course: DSC640 - Data Presentation and Visualization
Instructor: Catherine Williams

These two weeks we are going to be focused on histograms, box plots, and bullet charts and using various tools to create these visualizations. You must consolidate all the charts into ONE document with each chart labeled with the type of chart and technology - for example: Python - Bar Chart. Failure to label and consolidate the charts will resort in points being taken off or a 0 for the assignment.

Sample Datasets (click on the Downloads tab.)

You may also download them directly from this link: Exercise 6.2 Datasets (click the link to download a folder containing the datasets.)

You need to submit: 1 histogram, 1 box plot, 1 bullet chart, and 1 additional chart of your choice (can be an existing chart type we've already done, but it must be done in a new way or it can be an entirely new chart type) using Tableau or PowerBI

1 histogram, 1 box plot, 1 bullet chart, and 1 additional chart of your choice (can be an existing chart type we've already done, but it must be done in a new way or it can be an entirely new chart type) using Python

1 histogram, 1 box plot, 1 bullet chart, and 1 additional chart of your choice (can be an existing chart type we've already done, but it must be done in a new way or it can be an entirely new chart type) using R

1 histogram, 1 box plot, 1 bullet chart, and 1 additional chart of your choice using Python

In [4]:
```
##### Import common Data preparation & visualization libraries:
import numpy as np
import math
import matplotlib.pyplot as plt
import pandas as pd
import squarify
import seaborn as sns
import plotly.express as px
import matplotlib
```

In [5]:
```
## Ignore the warnings
```

```
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Read Input datasets

In [2]:

```
## Reading the birth rate dataset
birth_df = pd.read_csv('birth-rate.csv')
birth_df.head()
```

Out[2]:

	Country	1960	1961	1962	1963	1964	1965	1966	1967	1968	...	1999	2000	2001	2002	2003	2004	2005
0	Aruba	36.400	35.179	33.863	32.459	30.994	29.513	28.069	26.721	25.518	...	15.024	14.528	14.041	13.579	13.153	12.772	12.441
1	Afghanistan	52.201	52.206	52.208	52.204	52.192	52.168	52.130	52.076	52.006	...	51.229	50.903	50.486	49.984	49.416	48.803	48.177
2	Angola	54.432	54.394	54.317	54.199	54.040	53.836	53.585	53.296	52.984	...	48.662	48.355	48.005	47.545	46.936	46.184	45.330
3	Albania	40.886	40.312	39.604	38.792	37.913	37.008	36.112	35.245	34.421	...	17.713	16.850	16.081	15.444	14.962	14.644	14.485
4	Netherlands Antilles	32.321	30.987	29.618	28.229	26.849	25.518	24.280	23.173	22.230	...	15.809	15.412	15.096	14.824	14.565	14.309	14.051

5 rows × 50 columns

In [5]:

```
## Transposing birth dataset
birtht_df = pd.melt(birth_df, id_vars="Country", var_name="Year", value_name = 'BirthRate').fillna(0)
birtht_df["BirthRate_rnd"] = birtht_df["BirthRate"].apply(lambda x: math.ceil(x))
birtht_df.head(5)
```

Out[5]:

	Country	Year	BirthRate	BirthRate_rnd
0	Aruba	1960	36.400	37
1	Afghanistan	1960	52.201	53
2	Angola	1960	54.432	55
3	Albania	1960	40.886	41
4	Netherlands Antilles	1960	32.321	33

In [46]:
```
## Reading crime dataset
crime_df = pd.read_csv('crimeratesbystate-formatted.csv')
crime_df.head()
```

Out[46]:

	state	murder	forcible_rape	robbery	aggravated_assault	burglary	larceny_theft	motor_vehicle_theft
0	United States	5.6	31.7	140.7	291.1	726.7	2286.3	416.7
1	Alabama	8.2	34.3	141.4	247.8	953.8	2650.0	288.3
2	Alaska	4.8	81.1	80.9	465.1	622.5	2599.1	391.0
3	Arizona	7.5	33.8	144.4	327.4	948.4	2965.2	924.4
4	Arkansas	6.7	42.9	91.1	386.8	1084.6	2711.2	262.1

In [4]:
```
## Reading education dataset into dataframe
education_df = pd.read_csv('education.csv')
education_df.head()
```

Out[4]:

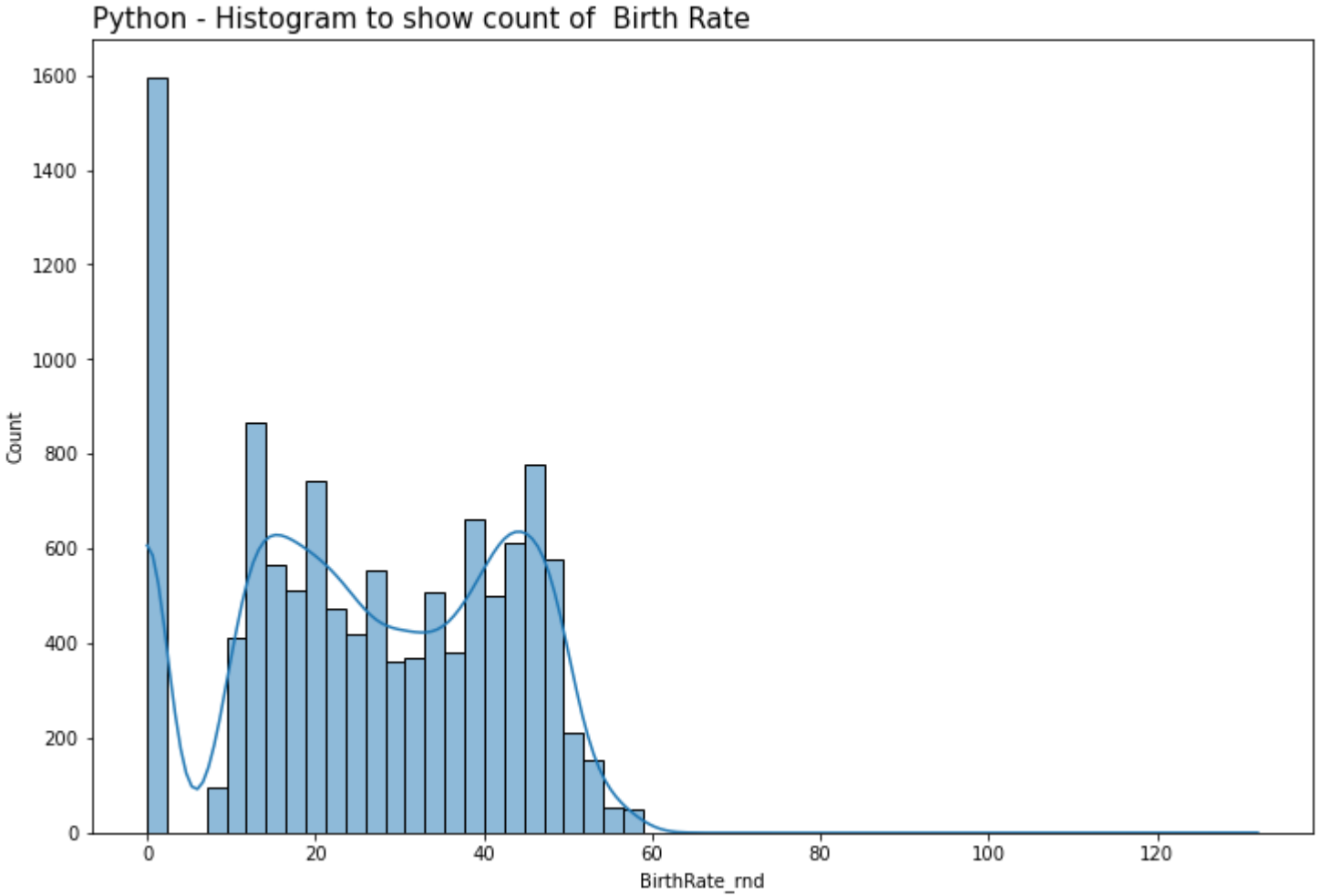
	state	reading	math	writing	percent_graduates_sat	pupil_staff_ratio	dropout_rate
0	United States	501	515	493	46	7.9	4.4
1	Alabama	557	552	549	7	6.7	2.3
2	Alaska	520	516	492	46	7.9	7.3
3	Arizona	516	521	497	26	10.4	7.6
4	Arkansas	572	572	556	5	6.8	4.6

In [54]:
```
# fix whitespaces from dataset
education_df = education_df.applymap(lambda x: x.strip() if type(x) is str else x)
crime_df = crime_df.applymap(lambda x: x.strip() if type(x) is str else x)
birth_df = birth_df.applymap(lambda x: x.strip() if type(x) is str else x)
```

1. Python - Histogram Plot

In [23]:
```
## Plotting histogram chart using sns histplot method
plt.figure(figsize=(12,8))
sns.histplot(data=birtht_df, x="BirthRate_rnd", kde=True)
```

```
plt.title("Python - Histogram to show count of  Birth Rate", fontsize = 15, loc = 'left')
plt.show()
```



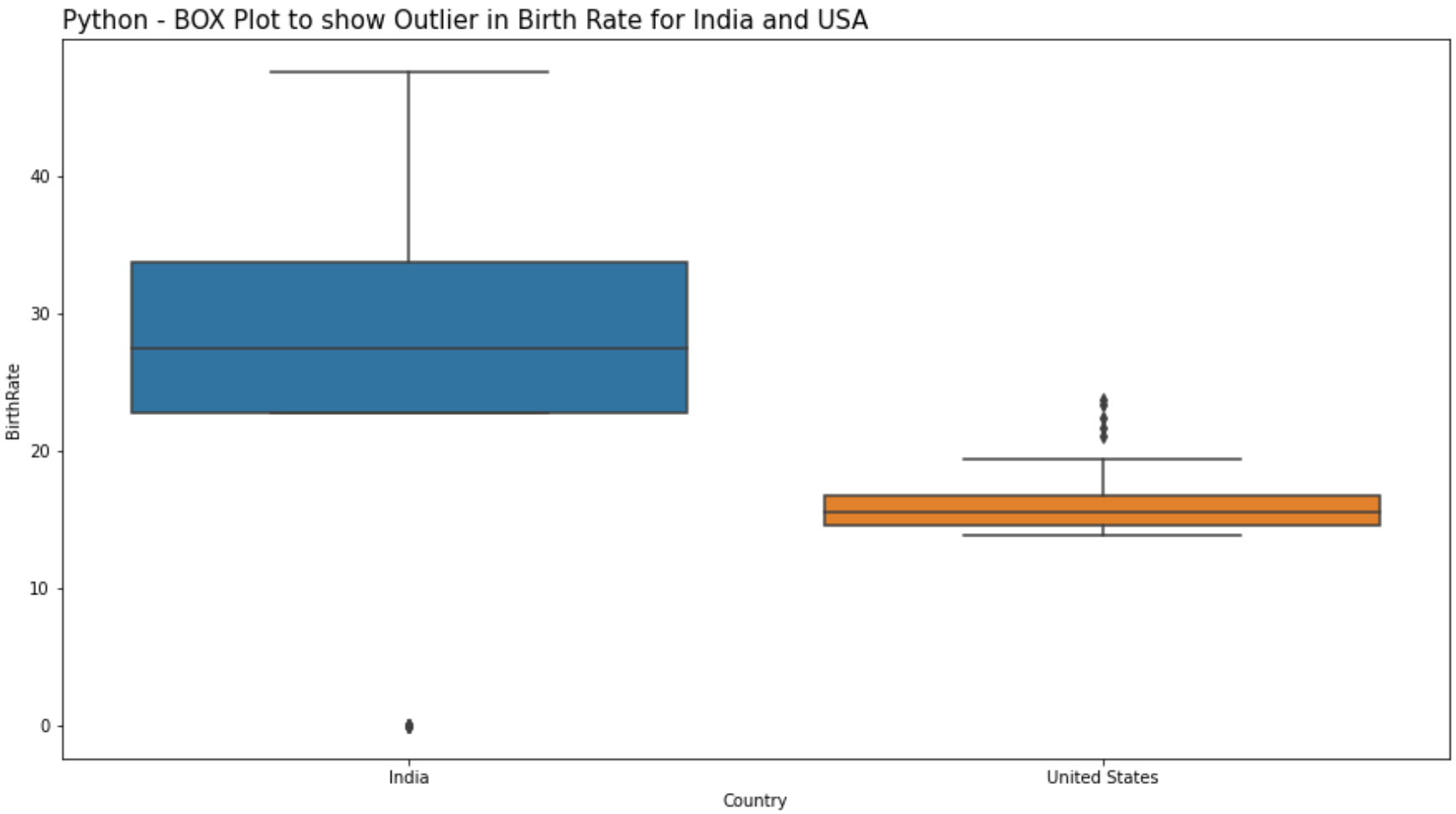
2. Python - Box Plot

Plotting Box Map for Birth Rate

In [27]:

```
plt.figure(figsize=(15,8))
birtht_box = birtht_df[(birtht_df["Country"]=="United States") | (birtht_df["Country"]=="India")]
sns.boxplot(x = birtht_box["Country"], y=birtht_box["BirthRate"])
```

```
plt.title("Python - BOX Plot to show Outlier in Birth Rate for India and USA", fontsize = 15, loc = 'left')
plt.show()
```



3. Python - Bullet Chart

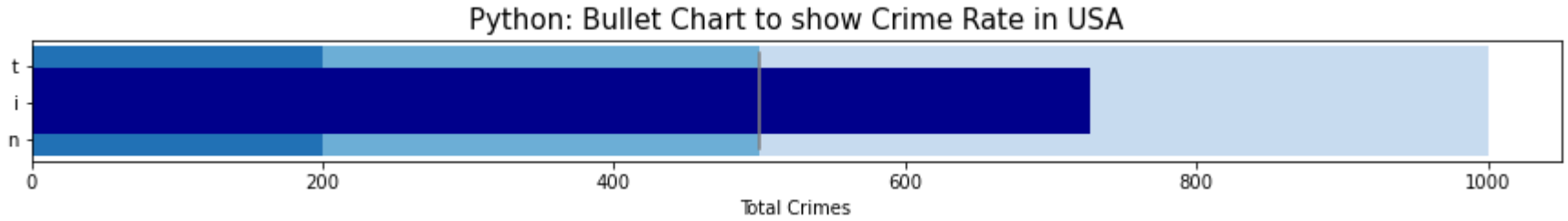
In [64]:
```
## Creating the dataset to be used for bullet chart
crime_bull = crime_df[crime_df["state"]=="United States"][["state","burglary"]]
crime_bull['target'] = 500
crime_bull_tuple = [tuple(x) for x in crime_bull.values][0]
crime_bull_tuple
```

Out[64]:
```
('United States', 726.7, 500)
```

In [65]:
```
# set parameter for bullet chart
limits = [200, 500, 1000]
palette = sns.color_palette("Blues_r", len(limits))
fig, ax = plt.subplots(figsize=(15,8))
ax.set_aspect('equal')
#ax.set_yticks([1])
ax.set_yticklabels(crime_bull_tuple[0])
prev_limit = 0
for idx, lim in enumerate(limits):
    ax.barh([1], lim-prev_limit, left=prev_limit, height=75, color=palette[idx])
    prev_limit = lim
    # draw the value we're measuring
ax.barh([1], crime_bull_tuple[1], color='darkblue', height=45)
ax.axvline(crime_bull_tuple[2], color="gray", ymin=0.10, ymax=0.9)
ax.set_title("Python: Bullet Chart to show Crime Rate in USA",fontsize=15)
ax.set_xlabel("Total Crimes")
```

```
<ipython-input-65-0f0cc7104dcb>:7: UserWarning: FixedFormatter should only be used together with FixedLocator
  ax.set_yticklabels(crime_bull_tuple[0])
```

Out[65]: Text(0.5, 0, 'Total Crimes')



4. Word Cloud

In [1]:
```
!pip install wordcloud
```

```
Collecting wordcloud
  Downloading wordcloud-1.8.2.2-cp38-cp38-win_amd64.whl (152 kB)
Requirement already satisfied: numpy>=1.6.1 in c:\users\kesavadithya\anaconda3\lib\site-packages (from wordcloud) (1.20.
1)
Requirement already satisfied: pillow in c:\users\kesavadithya\anaconda3\lib\site-packages (from wordcloud) (8.2.0)
Requirement already satisfied: matplotlib in c:\users\kesavadithya\anaconda3\lib\site-packages (from wordcloud) (3.3.4)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\kesavadithya\anaconda3\lib\site-packages (from matplotlib
->wordcloud) (2.8.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\kesavadithya\anaconda3\lib\site-packa
```

```
ges (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: cycler>=0.10 in c:\users\kesavadithya\anaconda3\lib\site-packages (from matplotlib->wordcl
oud) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\kesavadithya\anaconda3\lib\site-packages (from matplotlib->w
ordcloud) (1.3.1)
Requirement already satisfied: six in c:\users\kesavadithya\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib->w
ordcloud) (1.15.0)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.8.2.2
```

In [2]:
```
## Importing the lib for word cloud
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

In [16]:
```
## Creating the dataframe and printing few records
airtravel_df = pd.read_csv(r"airline_safety.txt", encoding ="latin-1")
airtravel_df.head()
```

Out[16]:

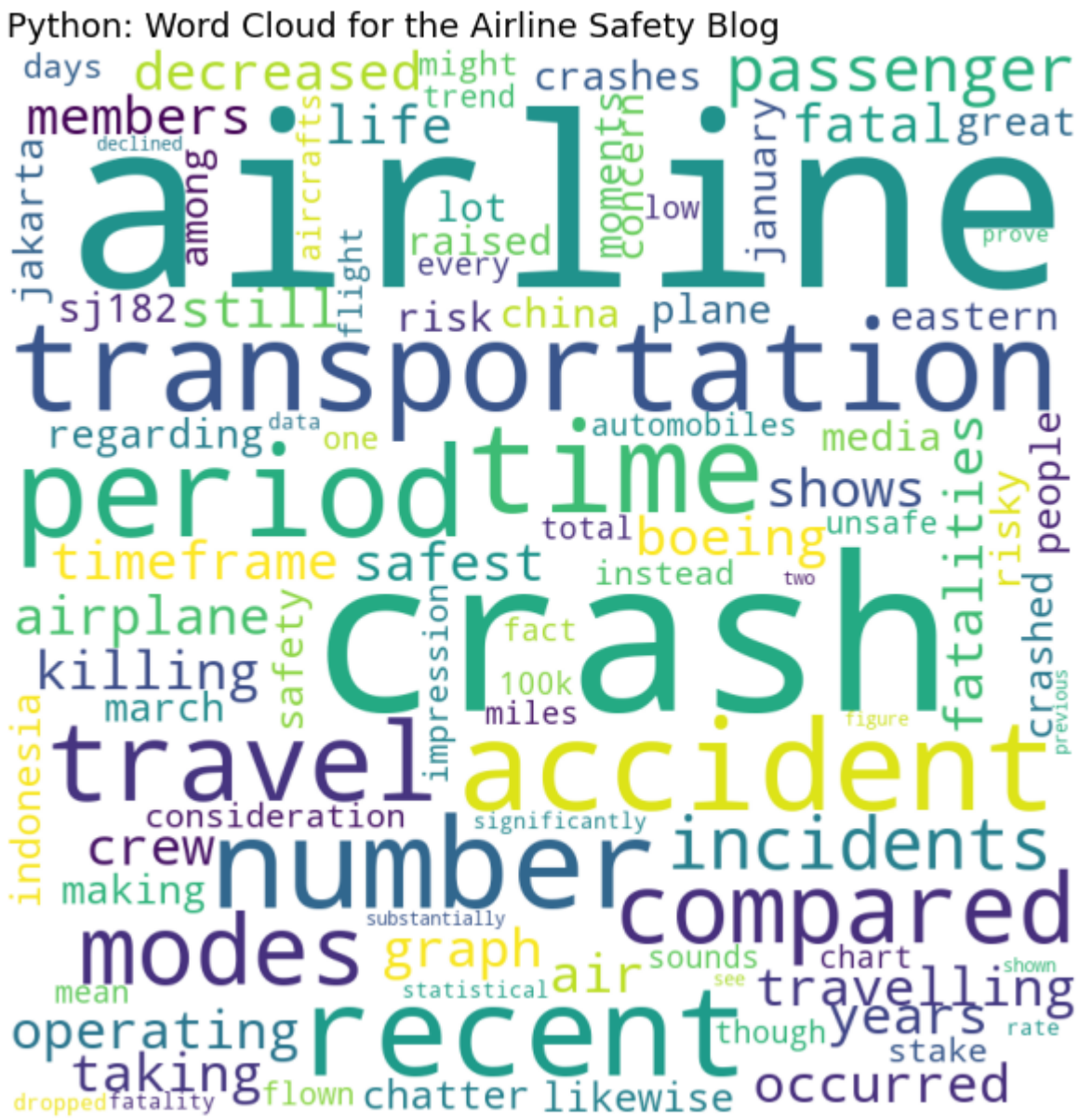
	content
0	Crash! Crash! Crash! Travelling in airplane is...
1	Air travel is risky compared to other modes of...
2	Likewise there has been a lot of chatter in th...
3	Recent incidents like a Boeing 737-800 plane o...
4	With these incidents in consideration you migh...

In [17]:
```
comment_words = ''
stopwords = set(STOPWORDS)
```

In [20]:
```
# iterate through the csv file
for val in airtravel_df.content:
    # typecaste each val to string
    val = str(val)
    # split the value
    tokens = val.split()
    # Converts each token into lowercase
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()
    comment_words += " ".join(tokens)+" "
```

In [26]:

```
wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)
plt.title("Python: Word Cloud for the Airline Safety Blog",fontsize = 18, loc = 'left')
plt.show()
```


Python: Word Cloud for the Airline Safety Blog
airline
transportation
crash
period
time
accident
travel
number
compared
incidents
modes
recent
passenger
fatal
great
members
life
decreased
crashes
might
trend
days
jakarta
among
declined
aircrafts
moments
concerns
january
lot
raised
every
low
prove
flight
sj182
still
risk
china
plane
eastern
regarding
data
one
automobiles
media
shows
total
boeing
unsafe
timeframe
safest
instead
two
fatalities
risky
people
crashed
airplane
killing
march
safety
impression
fact
100k
miles
figure
previous
indonesia
consideration
significantly
crew
making
substantially
graph
air
sounds
chart
shown
travelling
mean
statistical
see
operating
taking
years
though
rate
stake
occurred
likewise
chatter
flown
dropped
fatality

In [ ]:

Assignment_Week_9&10_Venkidusamy_KesavAdithya

Kesav Adithya Venkidusamy

2022/08/05

```
knitr::opts_chunk$set(echo = TRUE)

library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

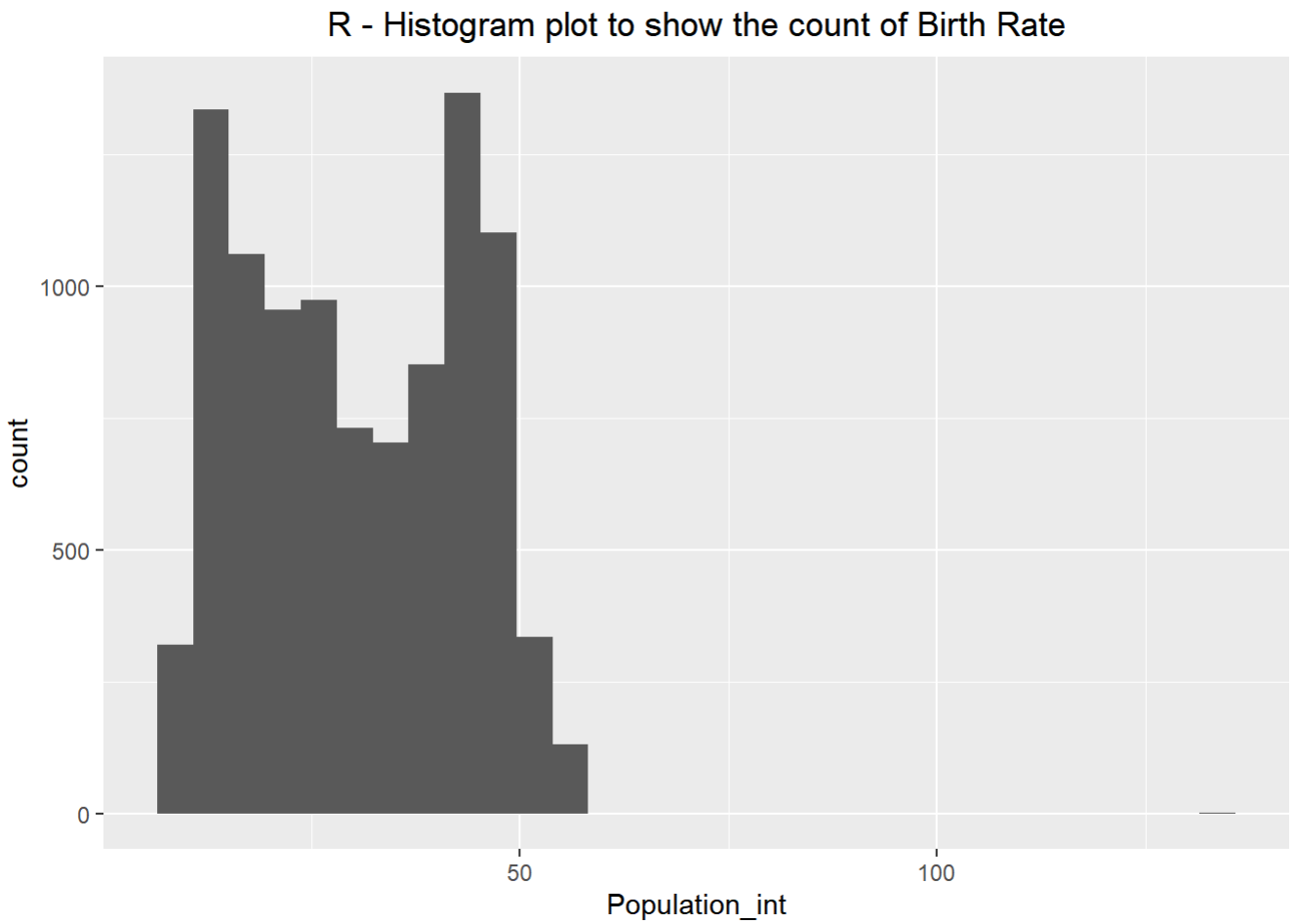
```
## The following object is masked from 'package:graphics':
##
##     layout
```

R: Histogram Plot

```
# Creating dataframe
birth_df <- read.csv("E:/Personal/Bellevue University/Course/github/dsc640/Week 9&10/birth-rate.csv")

# Format year column
colnames(birth_df) <- gsub("X", "", colnames(birth_df))

## Pivotting the birth dataframe
birtht_df <- reshape2::melt(birth_df, id=c("Country")) %>% dplyr::mutate("Country" = as.character(Country), "Year" = as.char
acter(variable), "Population" = value, "Population_int"=ceiling(value)) %>% dplyr::select(c("Country","Year","Population","P
opulation_int"))

ggplot(birtht_df, aes(x=Population_int)) + geom_histogram() + ggtitle("R - Histogram plot to show the count of Birth Rate")
 + theme(plot.title = element_text(hjust=0.5))
```

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

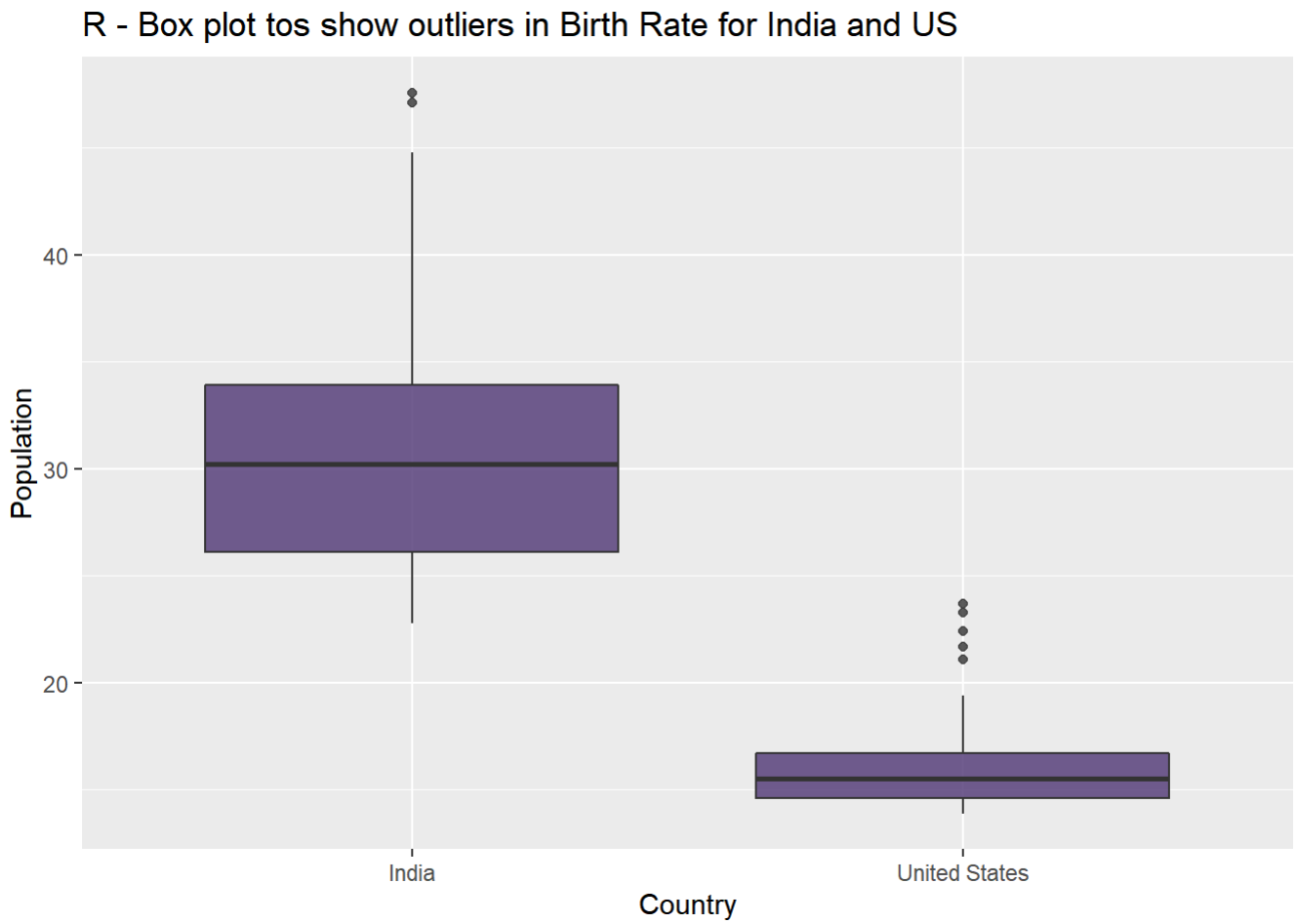
```
## Warning: Removed 1596 rows containing non-finite values (stat_bin).
```



R: Box Plot

```
## Create box plot

birth_box_df <- birtht_df %>% dplyr::filter(Country %in% c("United States","India"))

ggplot(birth_box_df, aes(x=Country, y=Population)) +
  geom_boxplot(fill="#4f3674", alpha=0.8) + ggtitle("R - Box plot tos show outliers in Birth Rate for India and US")
```

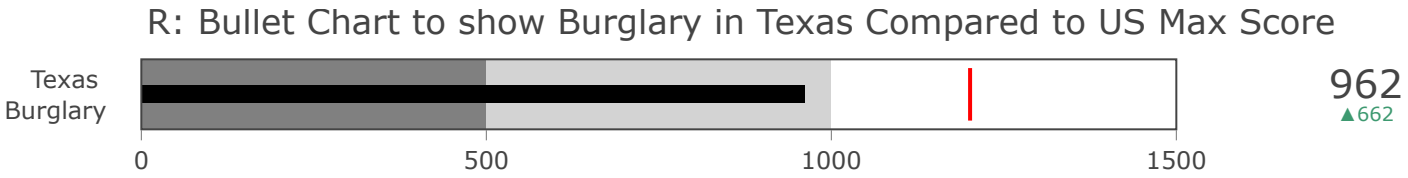
```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```



R: Bullet Chart

```
# Creating dataframe
crime_df <- read.csv("E:/Personal/Bellevue University/Course/github/dsc640/Week 9&10/crimeratesbystate-formatted.csv")

crime_bullet <- crime_df %>% dplyr::filter(stringr::str_trim(state, 'both') == "Texas") %>% dplyr::select(c(state, burglar
y))

maxburlgary <- max(crime_df$burglary)

fig <- plot_ly(
  type = "indicator",
  mode = "number+gauge+delta",
  value = crime_bullet$burglary,
  textposition = 'middle left',
  domain = list(x = c(0, 1), y= c(0, 1)),
  title = list(text = "Texas \nBurglary", font = list(size = 12)),
  delta = list(reference = 300),
  gauge = list(
    shape = "bullet",
    axis = list(range = list(NULL, 1500)),
    threshold = list(
      line = list(color = "red", width = 2),
      thickness = 0.75,
      value = maxburlgary),
    steps = list(
      list(range = c(0, 500), color = "gray"),
      list(range = c(500, 1000), color = "lightgray"),
      list(range = c(1000, 1500), color = "white")),
  bar = list(color = "black")),
  height = 100, width = 800)
fig <- fig %>%
  layout(margin = list(l= 100, r= 10))
fig <- fig %>%
  layout(title="R: Bullet Chart to show Burglary in Texas Compared to US Max Score", font = list(align = 'left'))

fig
```

```
## Warning: 'indicator' objects don't have these attributes: 'textposition'
## Valid attributes include:
## 'align', 'customdata', 'customdatasrc', 'delta', 'domain', 'gauge', 'ids', 'idssrc', 'legendgrouptitle', 'legendrank', 'm
eta', 'metasrc', 'mode', 'name', 'number', 'stream', 'title', 'transforms', 'type', 'uid', 'uirevision', 'value', 'visible',
'key', 'set', 'frame', 'transforms', '_isNestedKey', '_isSimpleKey', '_isGraticule', '_bbox'
```



R: Word Cloud

```
# Load libraries
library(tm)
```

```
## Warning: package 'tm' was built under R version 4.1.3
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.1.3
```

```
## Loading required package: RColorBrewer
```

```
library(SnowballC)
options(warn=-1)

# Read the data from file
airline_df <- read.csv("E:/Personal/Bellevue University/Course/github/dsc640/Week 9&10/airline_safety.txt")

# Create Corpus
corp <- VCorpus(VectorSource(airline_df))

# Clean up text data
corp <- tm_map(corp, removeNumbers)
corp <- tm_map(corp, removePunctuation)
corp <- tm_map(corp, stripWhitespace)
corp <- tm_map(corp, content_transformer(tolower))
corp <- tm_map(corp, removeWords, stopwords("english"))

# Create a document-term-matrix
dtm <- TermDocumentMatrix(corp)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix), decreasing = TRUE)
df <- data.frame(words=names(words), freq=words)

# Generate word cloud
wordcloud(words = df$words, freq=df$freq, min.freq = 1, max.words = 100, random.order = FALSE, colors = brewer.pal(8, "Dark
2"))
```

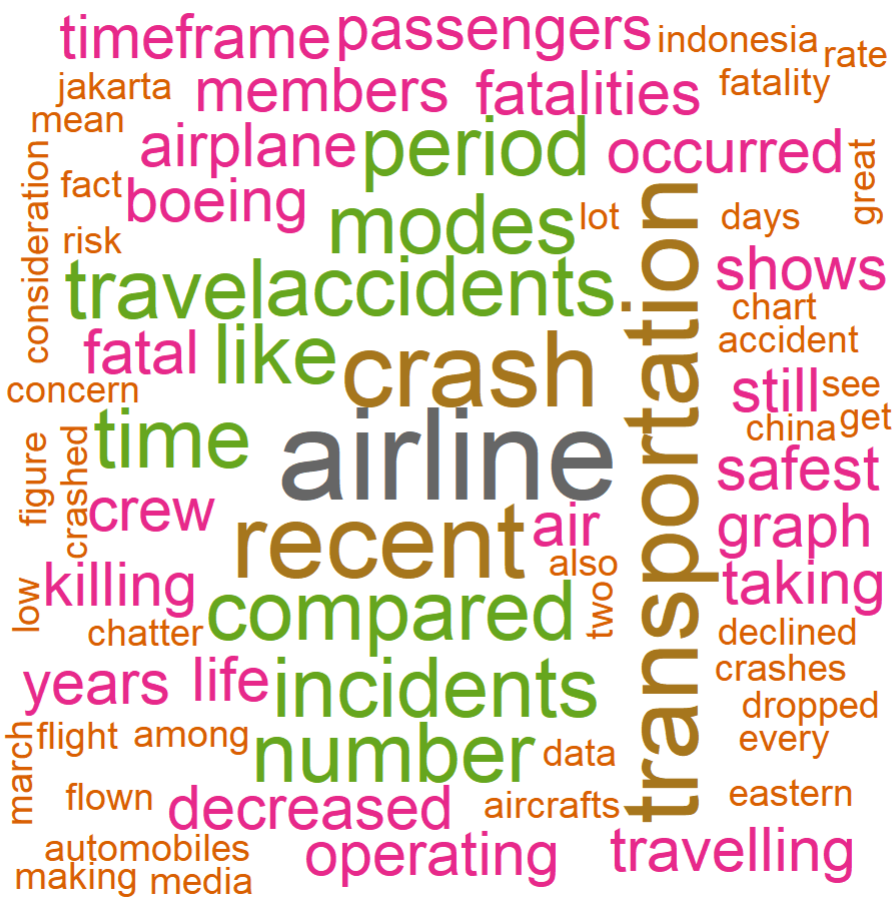
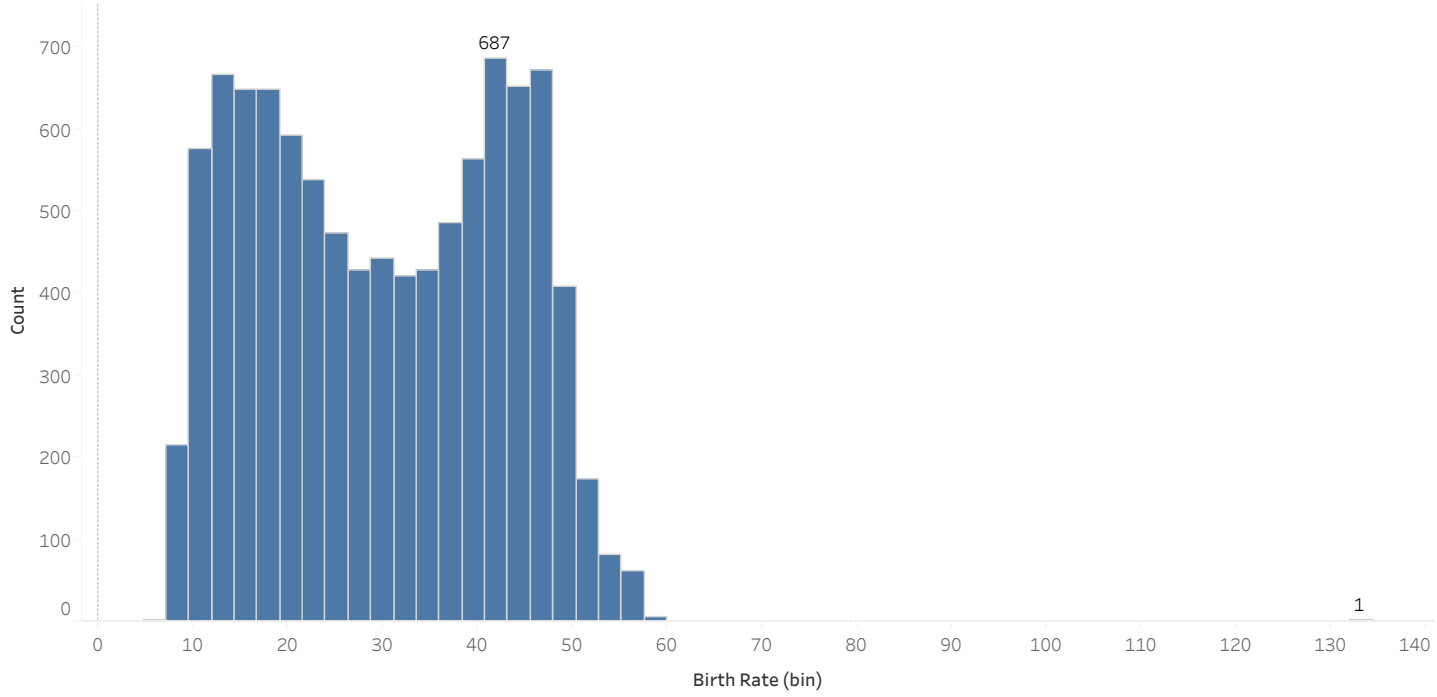
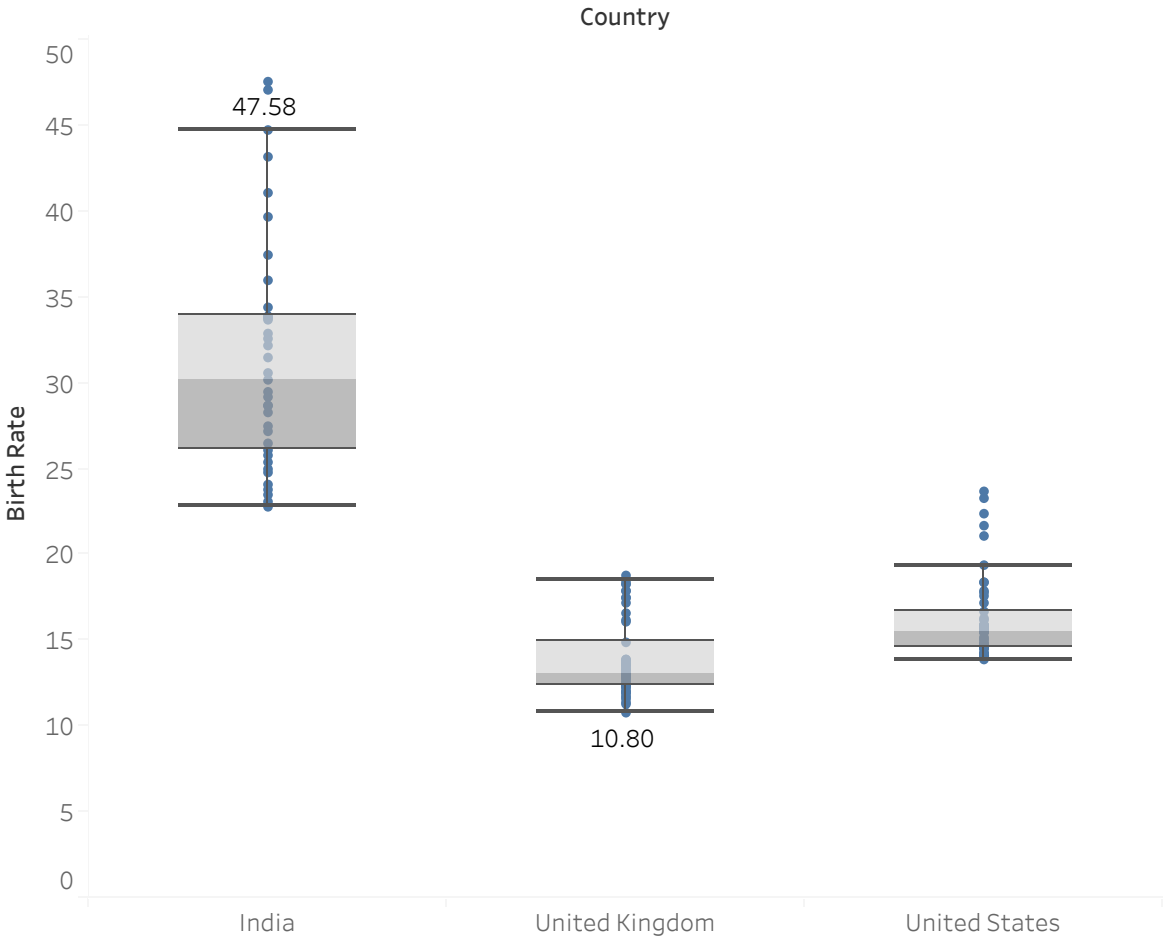

timeframe
passengers
indonesia
rate
jakarta
members
fatalities
fatality
mean
airplane
period
occurred
consideration
fact
boeing
great
risk
modes
lot
days
travelaccidents
shows
chart
fatal
like
crash
accident
concern
still
see
time
airline
china
get
figure
crashed
safest
crew
recent
air
graph
killing
also
taking
low
compared
two
chatter
declined
years
life
incidents
crashes
dropped
march
flight
among
number
data
every
flown
decreased
aircrafts
eastern
automobiles
operating
travelling
making
media
transportation

Tableau: Histogram to show the Birth Rate in US



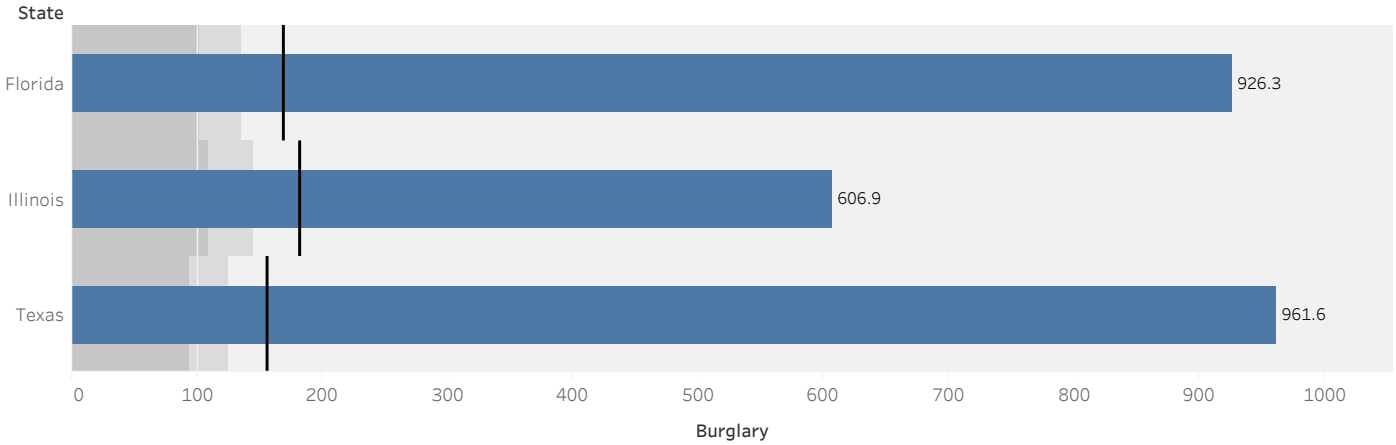
The trend of count of Birth Rate for Birth Rate (bin). The data is filtered on Birth Rate (bin), which excludes Null.

Tableau: Box Plot to show the Birth Rate for India, UK and US



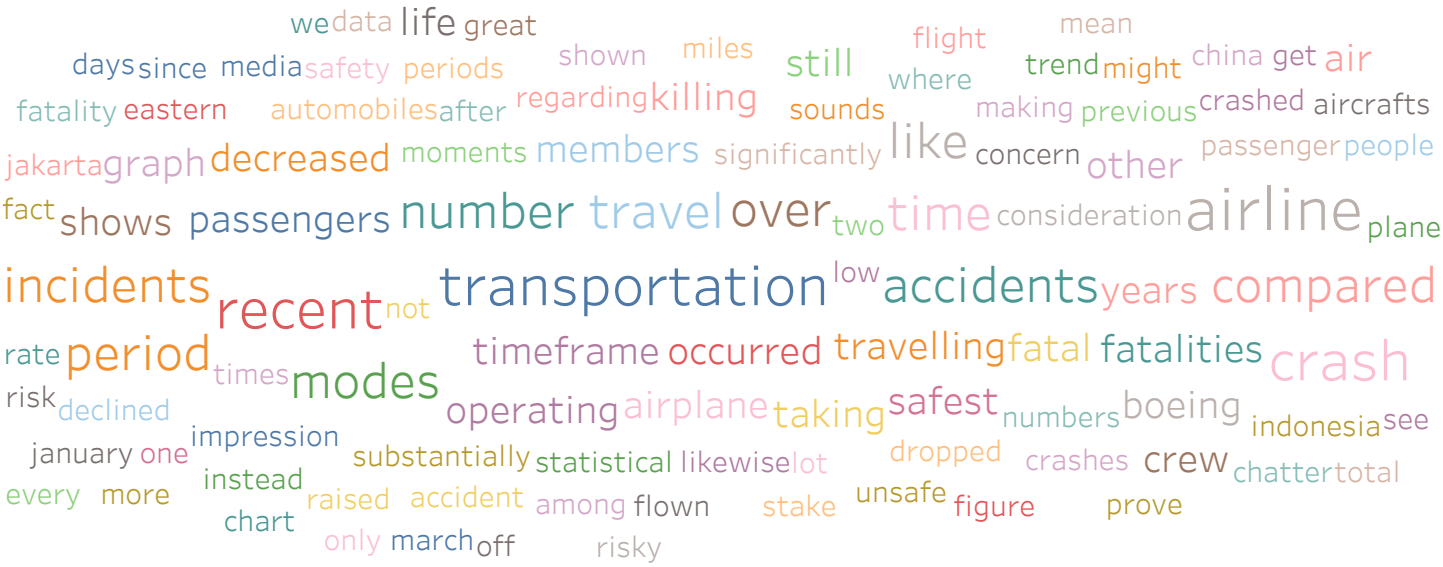
Sum of Birth Rate for each Country. Details are shown for Year. The view is filtered on Country and sum of Birth Rate. The Country filter keeps India, United Kingdom and United States. The sum of Birth Rate filter keeps non-Null values only.

Tableau: Bullet Chart to Show Burglary against Robbery in Florida/Texas/Illinois



Sum of Burglary for each State. The view is filtered on State, which keeps Florida, Illinois and Texas.

Tableau: Word Cloud for Airline Blog



Format_Content. Color shows details about Format_Content. Size shows count of Format_Content. The view is filtered on Format_Content, which keeps 109 of 145 members.