---
title: "BUDT758D Assignment 2:Charts"
author: "Emma Darkwa"
data: "03/08/2021"
output: pdf_document

Overview

An important element of data-driven decisions is the ability to visually communicate your data and interpretations. As "big data" and data analysis become more important yet more complex, it is even more important to understand the basic principles of data visualization.

Purpose

This assignment aligns with the second course objective: create visualizations using R.

Dataset and Submission Instructions

The dataset UMDBasketball2020.csv contains information of Maryland Terrapins school history from 1923 to 2019. The data was originally scraped from Sports-Reference.com. In this assignment, we will use this data set to study the team's overall wins and coaches' performance. A data dictionary can be found at the end of word document.

Visualization Guidelines

Make sure to change the axis titles, chart titles, colors, size, font, legend locations etc. if needed. Categories should be in a meaningful order, if appropriate. Also, format the grid lines and data series appropriately to enhance the clarity of the graph. Remember to write an informative title with some insights. Note that the outcome variable is typically on the y-axis (vertical axis).

You must turn in a well-formatted HTML or PDF output file. To do so, please click on the Knit button at the top with the wool ball icon, then choose to kint to HTML/PDF.

Q1. Explore the distribution of overall wins. (10 points)

a) Create a boxplot that examines the distribution of overall wins. (2 points)

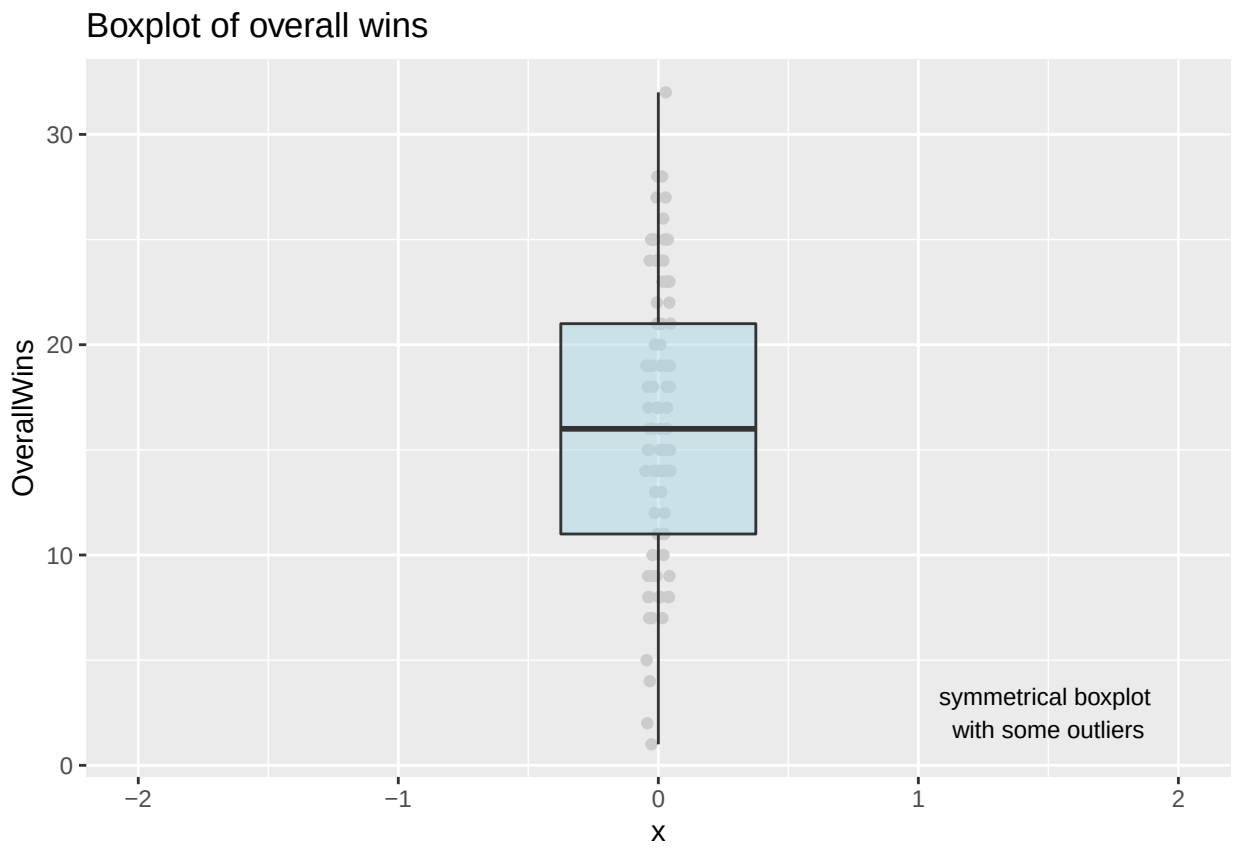
b) Add points of overall wins using the geom_jitter function. (2 points)

c) Add a general title to your chart such as "Boxplot of overall wins"; add a text box to describe your main finding in the chart using the annotate function. (3 points)

d) Improve your chart to make it clear and ready for presenting to your readers. (3 points)

(You only need to present a single chart with all the required information mentioned above)

```
ggp<-ggplot(data=umd_data,aes(x=0, y=OverallWins)) +
geom_jitter(width = 0.05, height = 0, color="grey80") +
geom_boxplot(fill="lightblue", alpha=0.5,outlier.color = NA) + xlim(-2,2)+
ggtitle('Boxplot of overall wins')

ggp + annotate("text", x=1.5,y=2.5,size=3,label="symmetrical boxplot \nwith some outliers")
```



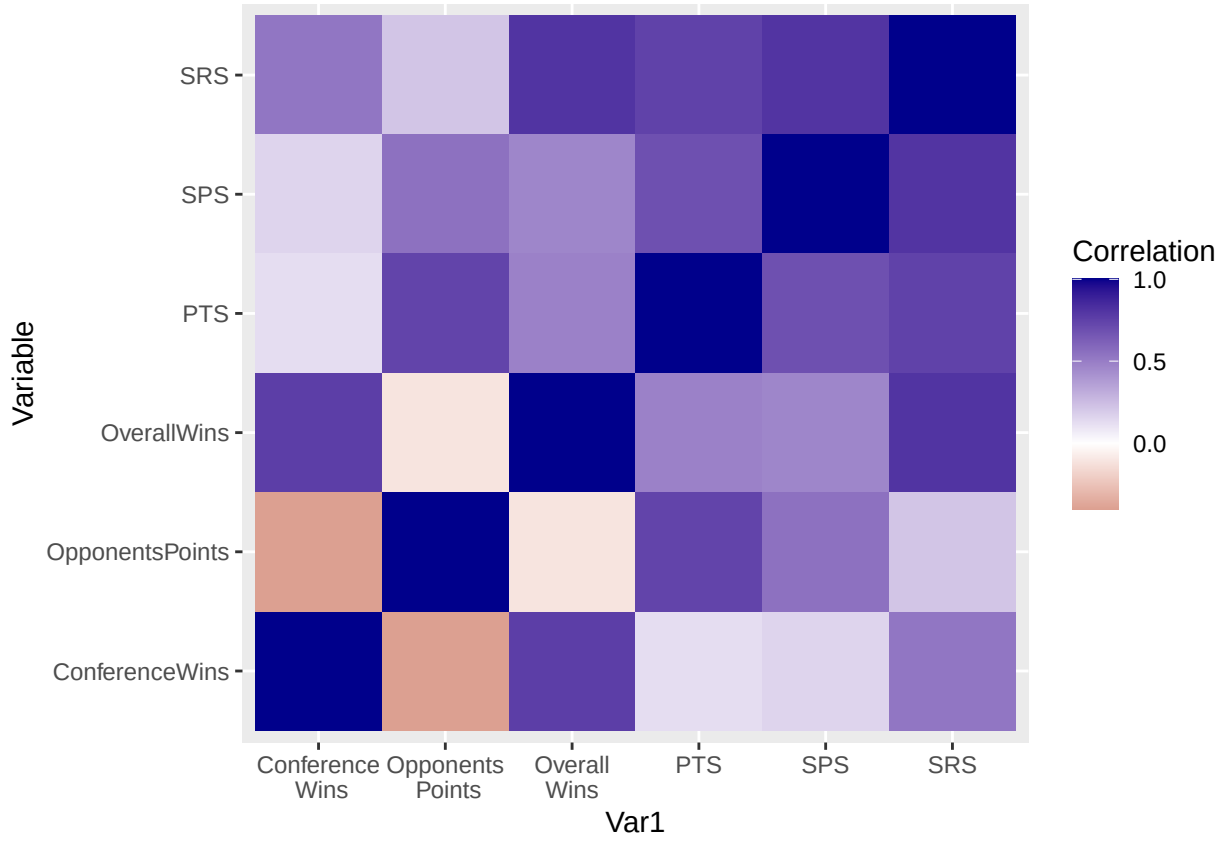
Q2. Explore the correlations between numeric variables. (10 points)

a) Create a correlations heat map for the following variables: OverallWins, ConferenceWins, SRS, SOS, PTS, and Opponents PTS. (3 points)

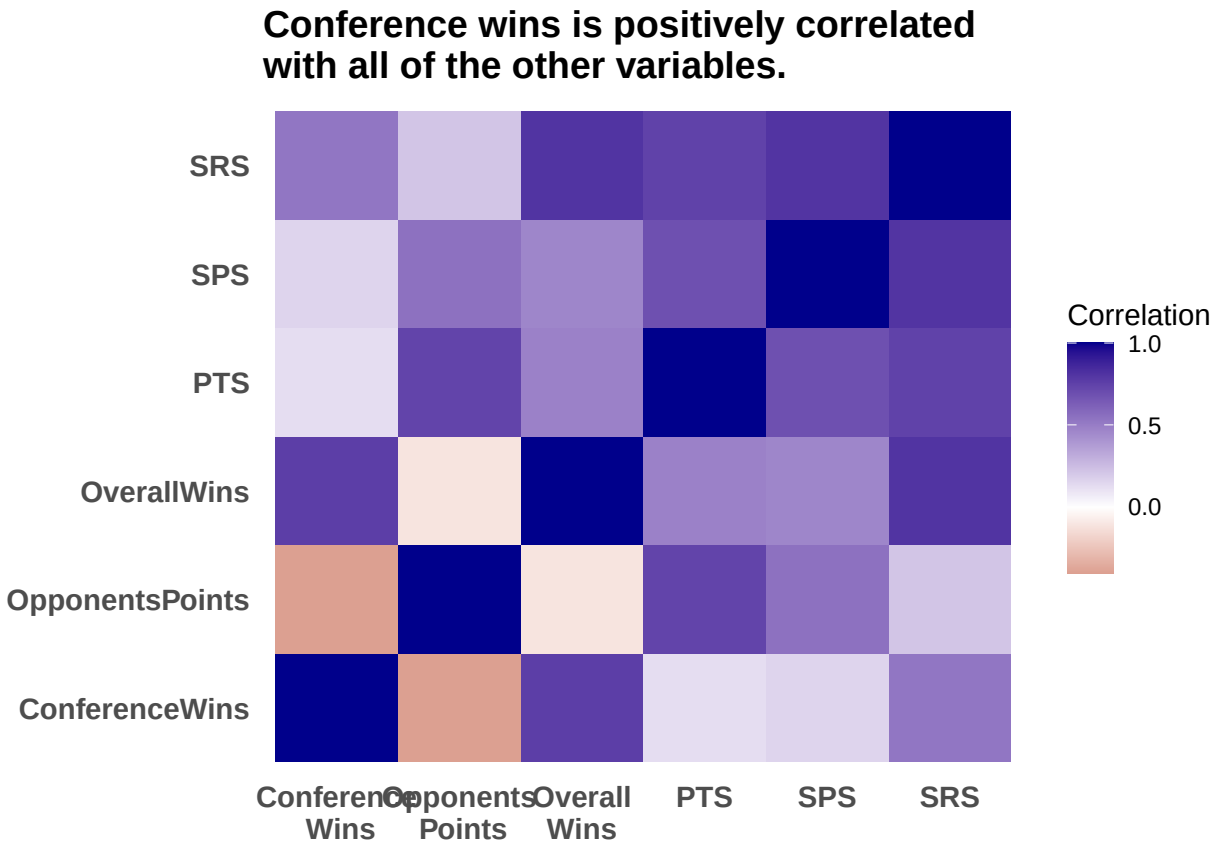
b) Improve your chart to make it clear and ready for presenting to your readers. (5 points)

```
data <-umd_data %>%
  select(OverallWins,ConferenceWins,SRS,SOS,PTS,OpponentsPTS)

data<-data[complete.cases(data),]
data1<-data %>%
  cor() %>%
  as.data.frame() %>%
  rownames_to_column(var = "Var1") %>%
  pivot_longer(OverallWins:OpponentsPTS, names_to ="Variable", values_to="corr")

# b) Make the chart
heat_map <- data1 %>%
  ggplot(aes(x=Var1, y=Variable, fill=corr)) +
  geom_tile()


umd_data_correlation1 <- heat_map +
  scale_x_discrete(labels= c("Conference\n Wins", "Opponents\n Points", "Overall\nWins","PTS", "SPS","SI
```

```
  scale_y_discrete(labels= c("ConferenceWins", "OpponentsPoints", "OverallWins","PTS", "SPS","SRS")) +
  scale_fill_gradient2(name="Correlation",midpoint = 0,
                       low="red4", mid="white", high = "blue4")
umd_data_correlation1
```




```
umd_data_correlation1 +
  labs(title = "Conference wins is positively correlated\nwith all of the other variables.") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title = element_blank(),
        axis.text = element_text(face="bold",size=11),
        plot.title = element_text(face="bold",size=14))
```



c) Which variables are positively correlated with overall wins? Which variable is most correlated with overall wins? (2 points)

Answer here: Conference wins is positively correlated with overall wins and the most correlated with overall wins

Q3. Explore the relationship between overall wins and conference wins. (12 points)

a) Create a scatter plot of the overall wins and conference wins; use different colors or shapes to denote difference conferences (ACC, Big Ten and Southern). (3 points)

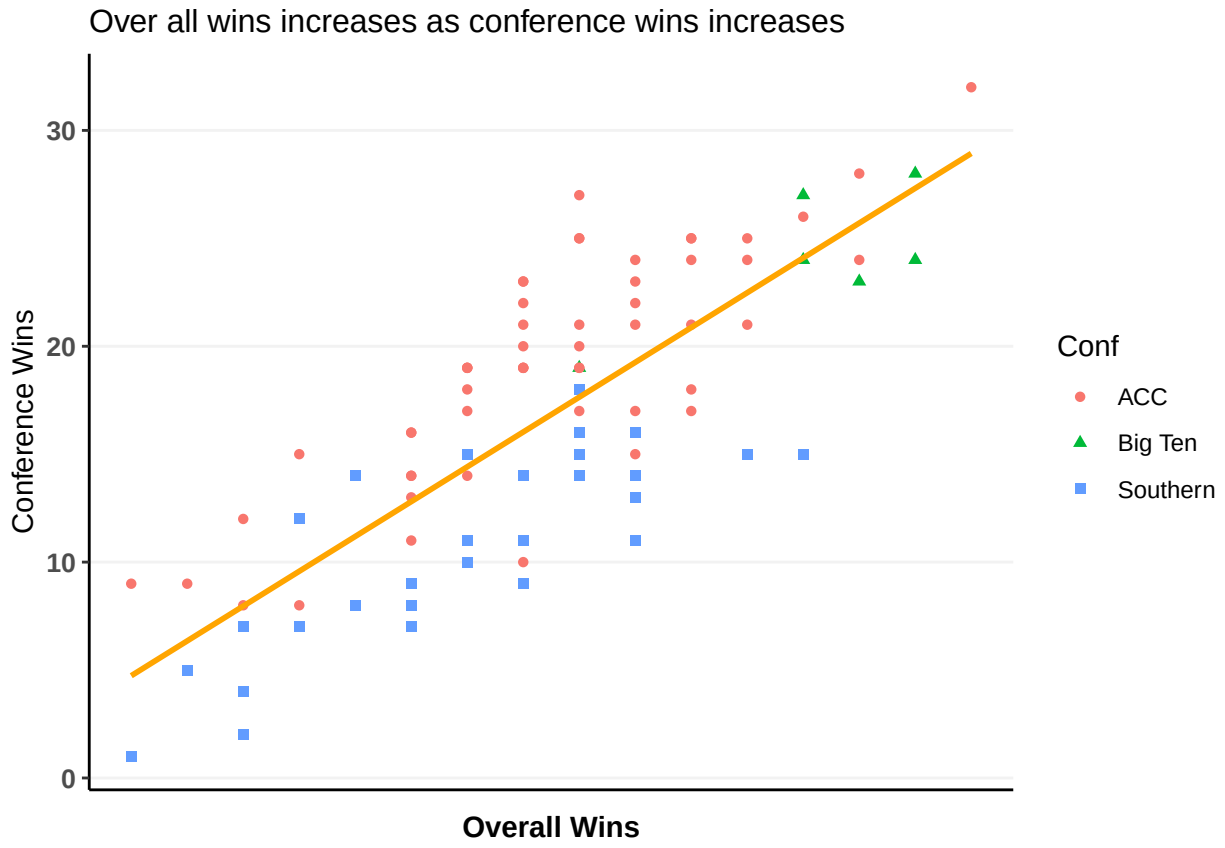
b) Add a single trend line to the chart. Hint: the mapping of color or shape needs to be created in the geom function instead of ggplot. (3 points)

c) Improve your chart to make it clear and ready for presenting to your readers. (3 points)

```
umd_data %>%
  ggplot(mapping=aes(x=ConferenceWins,y=OverallWins, colour=Conf)) +
  geom_point(aes(shape=Conf))+
  #geom_point(aes(color=factor(umd_data$Conf,pch=1,size=3)))+
    #scale_color_manual(values=c("lightblue","orange","red"))
  geom_smooth(method = "lm", se = FALSE, col='orange') + # se = FALSE will remove intervals around the
  labs(title="Over all wins increases as conference wins increases",
       x="Overall Wins", y="Conference Wins") +
  theme_classic() +
```

```
  theme(axis.title.x = element_text(face="bold",margin = margin(t = 10)), # x-axis title is too close t
        axis.text = element_text(face="bold",size=10),
        plot.caption = element_text(face="italic"),
        plot.title = element_text(size=12),
        panel.grid.minor = element_blank(), # remove minor grid lines
        panel.grid.major.y = element_line(color="grey95"),
        panel.grid.major.x = element_blank()) +
  scale_x_continuous(breaks = seq(1000,6000,1000))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



d) What pattern do you notice? (3 points)

Answer here: Positive linear relationship between the two numerical variables. Even though the Big Ten conference is not a lot it's in the highest part of the chart. There are also a lot of southern conference wins compared to big ten wins.

Q4. Explore the change of overall wins over years. (10 points)

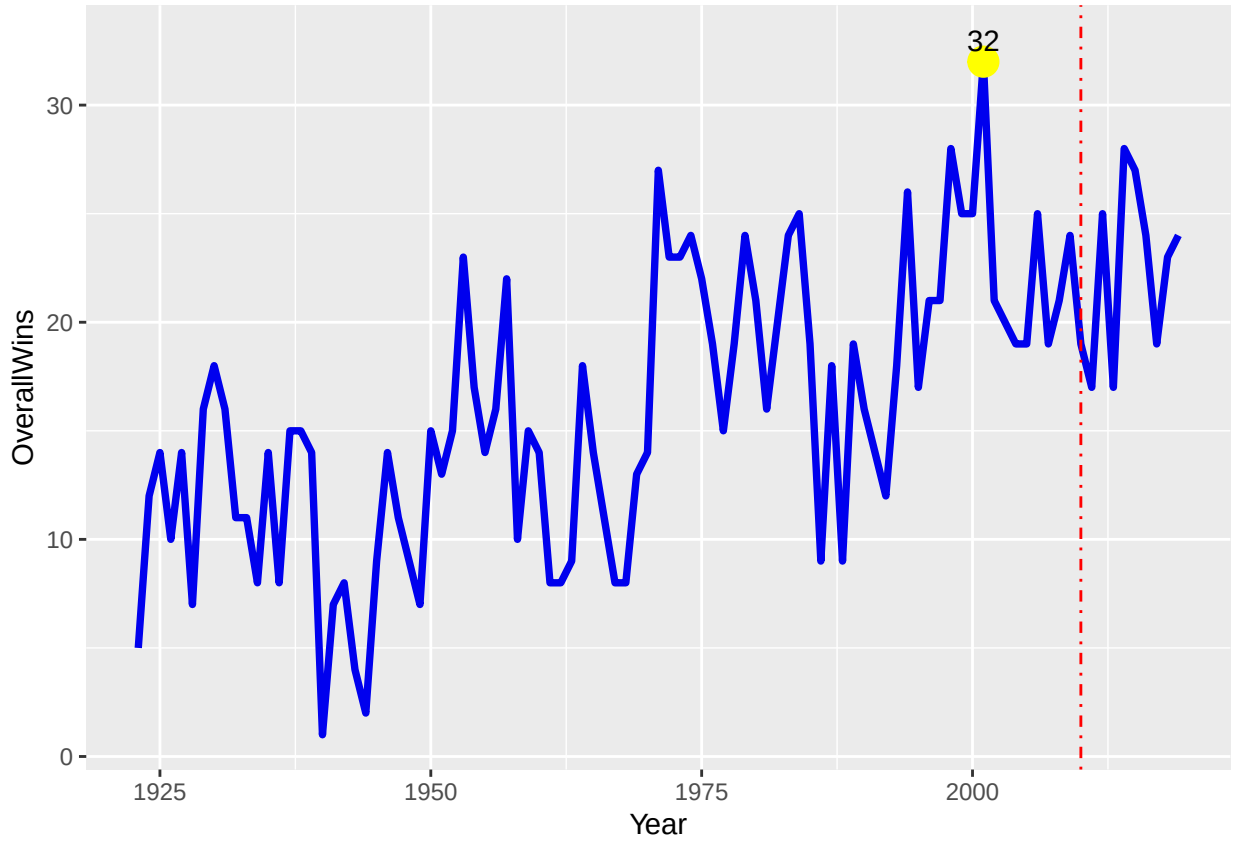
a) Create a line chart for the time series of overall wins. (2 points)

b) Highlight the point with the highest wins (Hint: you could use the geom_point function); add a data label above the point to show the value. (3 points)

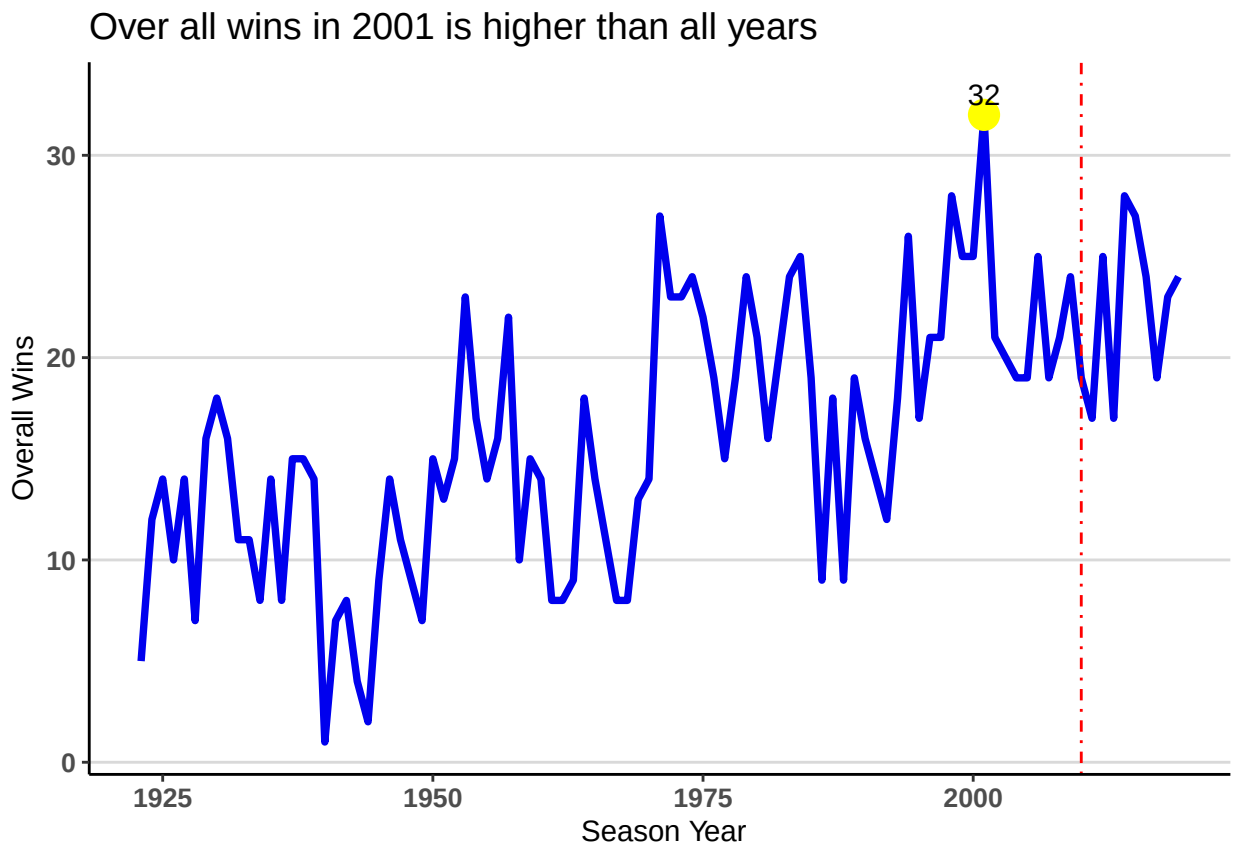
c) Add a vertical line at x = 2010. (2 points)

d) Improve your chart to make it clear and ready for presenting to your readers. (3 points)

```
umd_data_year <- umd_data


year_chart <- umd_data_year %>%
ggplot(aes(x=Year, y = OverallWins)) +
geom_line(col='blue2', lwd=1.3) +
geom_vline(xintercept = as.numeric("2010"),linetype=4,colour="red")+
geom_point(data=umd_data_year[which.max(umd_data_year$OverallWins),],color="yellow",size=5)+
geom_text(data = rbind(umd_data_year[which.max(umd_data_year$OverallWins),]),aes(Year,OverallWins+1,labe

year_chart
```




```
year_chart +
  labs(y = "Overall Wins",
       x = " Season Year",
       title = "Over all wins in 2001 is higher than all years",) +
  theme_classic() +
  theme(axis.text = element_text(face="bold",size=10),
        plot.caption = element_text(face="italic"),
        plot.title = element_text(size=14),
        panel.grid.major.y = element_line(color="grey85"))
```



Q5. Explore the number of seasons that each coach makes it to the NCAA tournament and the number of seasons he/she does not. (15 points)

(a) Create a stacked bar chart to show the number of seasons that each coach makes it to the NCAA tournament and the number of seasons he/she does not. Hint: create a new variable "NCAA" that indicates if the variable "NCAA Tournament" is NA or not. (4 points)

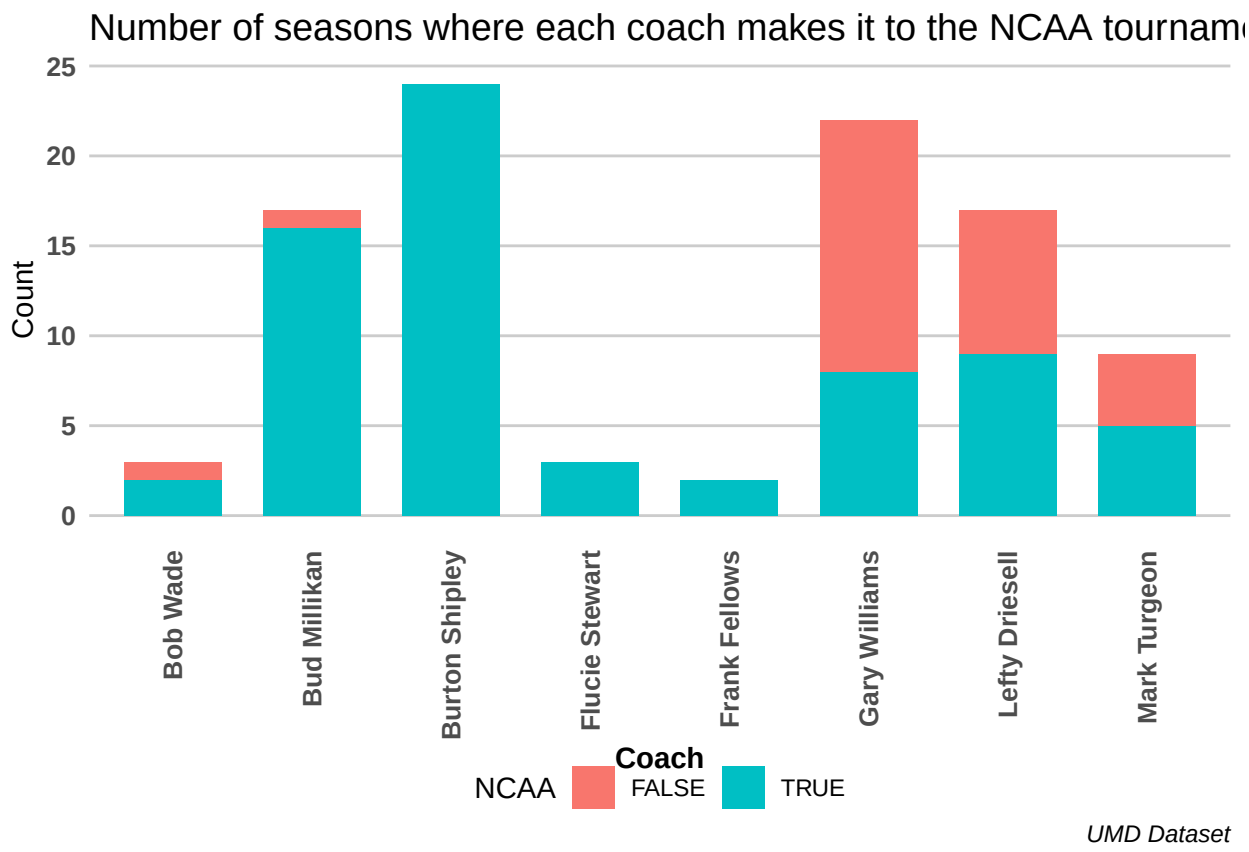
(b) Order the coaches based on their first year of serving as the coach at UMD. Hint: group by coach and then create a new variable first_year which is the minimum of the "Year" variable. Reorder "Coach" based on this variable. Then, group by Coach and NCAA and calculate the number of seasons within each group. (4 points)

(c) Improve your chart to make it clear and ready for presenting to your readers. (4 points)

```
umd_data$NCAA <- is.na(umd_data$`NCAA Tournament`)

umd_data$first_year <-umd_data%>%
  group_by(umd_data$Coach) %>%
  mutate(first_year =min(umd_data$Year)) %>%
  ungroup() %>%
  mutate (first_year =fct_reorder (Coach, first_year, .desc =FALSE))

umd_data1 <-  umd_data %>%
  group_by(Coach,NCAA) %>%
  summarize(count = n())
```

```
## `summarise()` has grouped output by 'Coach'. You can override using the `.groups` argument.
```

```r
# Create the chart
stackedbar <- umd_data1 %>%
  ggplot(aes(x=Coach, y=count)) +
  geom_col(aes(fill=NCAA),width=0.7) +

  # Add labels
  labs(y = "Count",
       x = "Coach",
       caption = "UMD Dataset",
       title = "Number of seasons where each coach makes it to the NCAA tournament") +
  # Set up the theme
  theme_minimal() +
  theme(axis.title.x = element_text(face="bold"),
        axis.text = element_text(face="bold",size=10),
        legend.position="bottom",
        legend.margin = margin(t=-15),
        plot.caption = element_text(face="italic"),
        plot.title = element_text(size=14),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        panel.grid.minor = element_blank(),
        panel.grid.major.y = element_line(color="grey80"),
        panel.grid.major.x = element_blank())
stackedbar
```



(d) Which coach is the best in terms of the number of seasons that he/she makes it to the NCAA? (3 points)

Answer here: Burton Shipley

Quality of code (3 points)

Your code needs to be clean, clear, and easy to follow.