DATA 607 Project 2 Part 3

Subhalaxmi Rout

03/08/2020

State Marriage Rates

Discussion thread created by : Gabriel Abreu

1. Introduction

These gives state marriage rates breaking down the data into regions and years. We can group the data by census region or census division. Then organize the rates according to year, changing it from wide data to long data.

URL: Dataset link

2.Load library

```
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("ggplot2")
#install.packages("DT")
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(tidyr)
library(ggplot2)
library(DT)
```

3. Data load and cleaning

Data is stored in the **Github** and loaded data from Github to Rstudio using **read.csv()** method.

```
# read csv file data in a variable using  read.csv
data <- read.csv('https://raw.githubusercontent.com/SubhalaxmiRout002/Data-607-Project-2-Dataset-1/mast

# convert data to data frame
data <- data.frame(data)

# display data using datatable
datatable(data,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	state	census_division	census_region	X2016	X2015	X2014	X2013	X2012	X2011	X
1	Alabama	East South Central	South	7.1478	7.3515	7.8068	7.8178	8.2	8.4	
2	Alaska	Pacific	West	7.1034	7.4076	7.5088	7.2939	7.2	7.8	
3	Arizona	Mountain	West	5.9305	5.9225	5.7804	5.4011	5.6	5.7	
4	Arkansas	West South Central	South	9.861	10.0403	10.112	9.7511	10.9	10.4	
5	California	Pacific	West	6.4636	6.185	6.4415	6.4605	6	5.8	
6	Colorado	Mountain	West	7.4254	6.7918	7.0616	6.4527	6.8	7	
7	Connecticut	New England	Northeast	5.6179	5.292	5.3688	5.021	5.2	5.5	
8	Delaware	South Atlantic	South	5.6131	5.7129	6.0228	6.572	5.8	5.2	
9	District of Columbia			8.1492	8.2204	11.8213	10.7913	8.4	8.7	
10	Florida	South Atlantic	South	8.126	8.2344	7.3014	7.0096	7.2	7.4	

Showing 1 to 10 of 51 entries Previous 1 2 3 4 5 6 Next

3.1 Gather year from 1990 to 2016

This dataset year has given from 1990 to 2016. Each year mentioned as a column. Using **tidyr** convert these columns to `Year` column.

```
# using gather() convert column to row
data <- data %>% gather(Year, Marriage_Rate, X2016:X1990, na.rm = TRUE)

# remove "X" from the year
data$Year  <- sub('X','',data$Year)

# arrange Marriage_Rate by desc order
data <- data %>% arrange(desc(Marriage_Rate))

# round Marriage Rate till 2 decimal
data$Marriage_Rate  <- round(data$Marriage_Rate,2)

# display data using datatable
datatable(data,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	state	census_division	census_region	Year	Marriage_Rate
1	Nevada	Mountain	West	1990	99
2	Nevada	Mountain	West	1995	85.2
3	Nevada	Mountain	West	1999	82.3
4	Nevada	Mountain	West	2000	72.2
5	Nevada	Mountain	West	2001	69.6
6	Nevada	Mountain	West	2002	67.4
7	Nevada	Mountain	West	2003	63.9
8	Nevada	Mountain	West	2004	62.1
9	Nevada	Mountain	West	2005	57.4
10	Nevada	Mountain	West	2006	52.1

Showing 1 to 10 of 1,013 entries Previous 1 2 3 4 5 … 102 Next

3.2 Rename column name

```
# rename census_divison and census_region i.e "Division" and "Region"
data <- data %>% rename( Division = census_division,Region = census_region)

# replace null value to NA
data$Region[data$Region == ""] <- NA

# display data using datatable
datatable(data,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	state	Division	Region	Year	Marriage_Rate
1	Nevada	Mountain	West	1990	99
2	Nevada	Mountain	West	1995	85.2
3	Nevada	Mountain	West	1999	82.3
4	Nevada	Mountain	West	2000	72.2
5	Nevada	Mountain	West	2001	69.6
6	Nevada	Mountain	West	2002	67.4
7	Nevada	Mountain	West	2003	63.9
8	Nevada	Mountain	West	2004	62.1
9	Nevada	Mountain	West	2005	57.4
10	Nevada	Mountain	West	2006	52.1

Showing 1 to 10 of 1,013 entries Previous 1 2 3 4 5 … 102 Next

3.3 Region wise Marriage rate

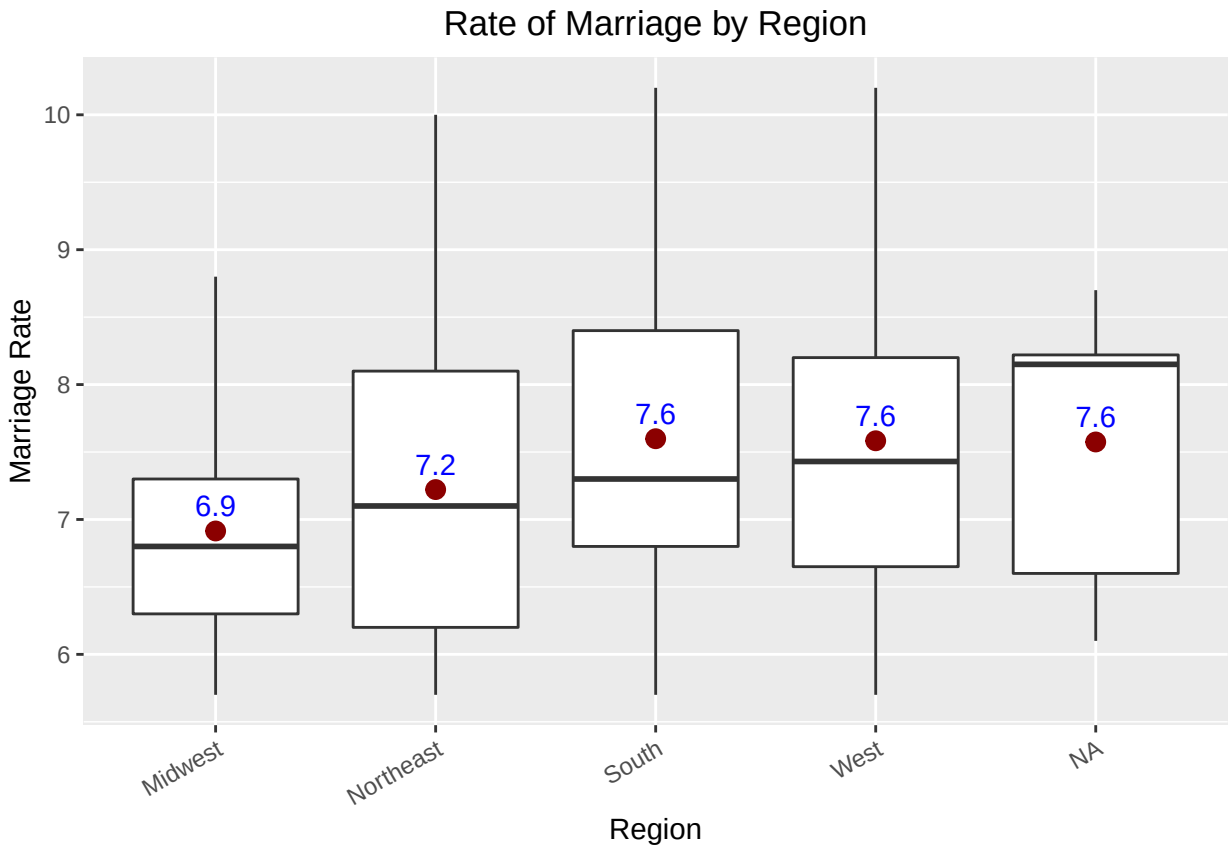
Apply group by on Region and plot the graph using region wise marriage rate.

```r
#load data in a data frame
data1 <- data.frame(data)


#apply groupby on Region
data1 <- data1 %>% group_by(Region, Year) %>% select(Region,Year, Marriage_Rate)

#  Boxplot to analyse marriage rate
ggplot(data1,aes(x = Region, y = Marriage_Rate)) +
  geom_boxplot(outlier.shape = NA) +
scale_y_continuous(limits = quantile(data1$Marriage_Rate, c(0.1, 0.9))) +
stat_summary(fun=mean, colour="darkred", geom="point", size=3,show.legend = FALSE) +
stat_summary(fun=mean, colour="blue", geom="text", show.legend = FALSE,
               vjust=-0.7, aes( label=round(..y.., digits=1))) +
  xlab("Region") + ylab("Marriage Rate") +
theme(axis.text.x=element_text(angle=30,hjust=1),plot.title = element_text(hjust = 0.5)) +
  ggtitle("Rate of Marriage by Region")
```

```
## Warning: Removed 186 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 186 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 186 rows containing non-finite values (stat_summary).
```



```r
# get all unique Division
Regions = unique(data1$Region)
Regions
```

```
## [1] "West"      "South"     NA          "Midwest"   "Northeast"
```

```r
# summary for all Region
data4 <- data1 %>%
  group_by(Region) %>%
  summarize(Min. = min(Marriage_Rate),
            "1st Qu." = round(quantile(Marriage_Rate, 0.25),2),
            Median = round(median(Marriage_Rate),2),
            Mean = round(mean(Marriage_Rate),2),
            "3rd Qu." = round(quantile(Marriage_Rate, 0.75),2),
            Max. = max(Marriage_Rate)
            )
# display summary using datatable
datatable(data4,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	Region	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	Midwest	5.22	6.1	6.7	6.79	7.2	11.1
2	Northeast	4.8	5.76	6.7	6.87	7.8	10.9
3	South	4.8	6.82	7.42	8.13	9	15.9
4	West	4	6.9	7.8	12.16	9.83	99
5		4	4.85	6.15	6.61	8.21	11.82

Showing 1 to 5 of 5 entries Previous 1 Next

3.4 Division wise Marriage rate

Apply group by on Division and plot the graph using region wise marriage rate.

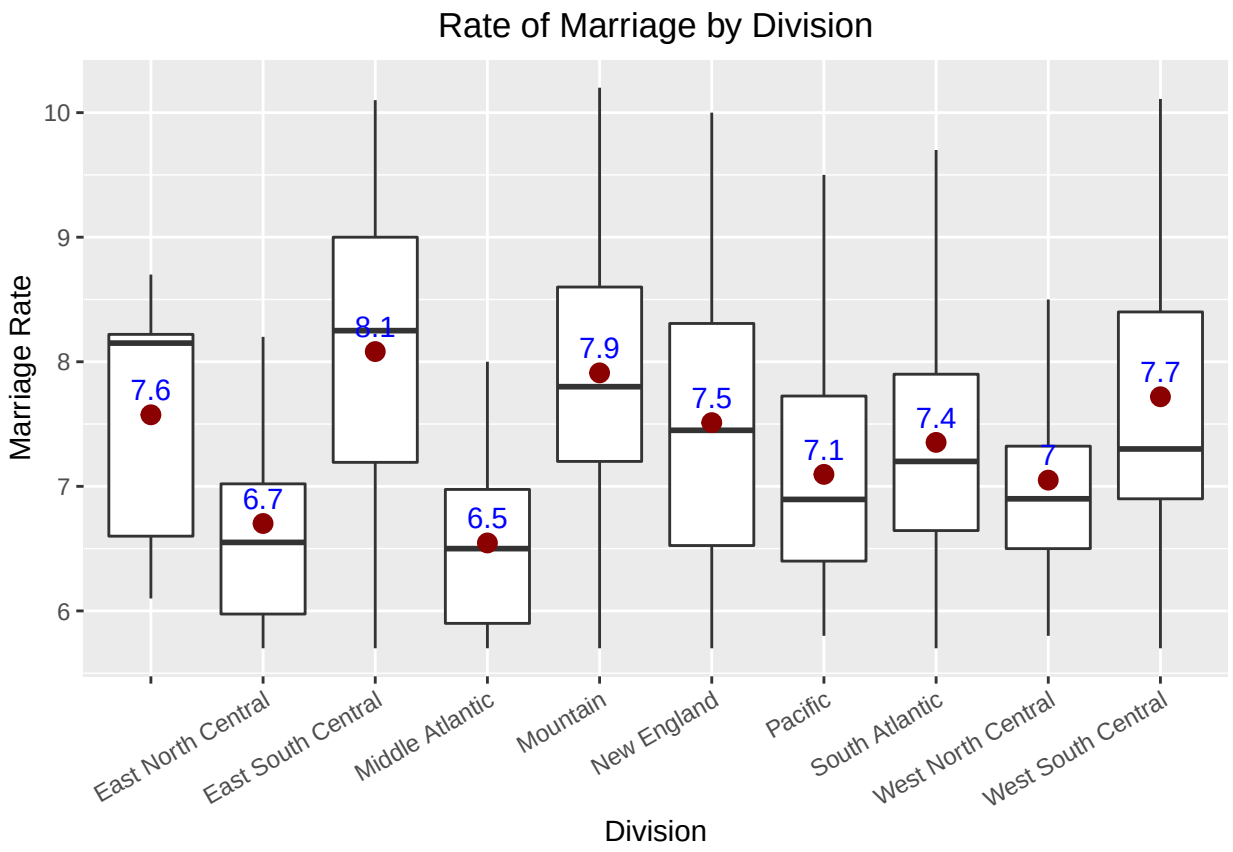
```r
#load data in a data frame
data2 <- data.frame(data)


#apply groupby on Region
data2 <- data2 %>% group_by(Division, Year) %>% select(Division,Year, Marriage_Rate)

# replace null value to NA
data$Division[data$Division == ""] <- NA

# plot line chart to analyse trend
ggplot(data2,aes(x = Division, y = Marriage_Rate)) +
  geom_boxplot(outlier.shape = NA) +
  scale_y_continuous(limits = quantile(data2$Marriage_Rate, c(0.1, 0.9))) +
stat_summary(fun=mean, colour="darkred", geom="point", size=3,show.legend = FALSE) +
stat_summary(fun=mean, colour="blue", geom="text", show.legend = FALSE,
               vjust=-0.7, aes( label=round(..y.., digits=1))) +
```

```
xlab("Division") + ylab("Marriage Rate") +
theme(axis.text.x=element_text(angle=30,hjust=1),plot.title = element_text(hjust = 0.5)) +
  ggtitle("Rate of Marriage by Division")
```

```
## Warning: Removed 186 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 186 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 186 rows containing non-finite values (stat_summary).
```




```
# get all unique Division
Divisions = unique(data2$Division)
Divisions
```

```
##  [1] "Mountain"           "Pacific"            "South Atlantic"
##  [4] "East South Central" "West South Central" ""
##  [7] "West North Central" "New England"        "East North Central"
## [10] "Middle Atlantic"
```

```
# summary for all division
data3 <- data2 %>%
  group_by(Division) %>%
  summarize(Min. = min(Marriage_Rate),
            "1st Qu." = round(quantile(Marriage_Rate, 0.25),2),
```

```
            Median = round(median(Marriage_Rate),2),
            Mean = round(mean(Marriage_Rate),2),
            "3rd Qu." = round(quantile(Marriage_Rate, 0.75),2),
            Max. = max(Marriage_Rate)
            )
# display data using datatable
datatable(data3,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	Division	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1		4	4.85	6.15	6.61	8.21	11.82
2	East North Central	5.22	5.8	6.25	6.5	7	9.6
3	East South Central	4.8	7.19	8.43	8.74	9.65	15.5
4	Middle Atlantic	4.8	5.58	5.9	6.18	6.8	8.6
5	Mountain	4	7.32	8.08	13.86	10	99
6	New England	5.02	5.9	7.14	7.22	8.3	10.9
7	Pacific	5.8	6.46	7.2	9.44	8.67	22.6
8	South Atlantic	5.2	6.6	7.2	7.45	8.12	15.9
9	West North Central	5.3	6.49	6.82	7	7.3	11.1
10	West South Central	5.7	7.09	8	8.89	10.07	15.4

Showing 1 to 10 of 10 entries Previous 1 Next

4. Analysis

We will analyze, yealy Marriage rate over Division and Region. We will look the trend, how the trend is changing over 26 years.

4.1 Yealy Average Marriage Rate over Division

```
# group by data by Division and Year
data5 <- data.frame(data) %>% group_by(Division, Year) %>% summarise(mean(Marriage_Rate)) %>% filter(!is

# rename Avg Marriage Rate column
data5 <- data5 %>% rename( Avg_Marriage_Rate = `mean(Marriage_Rate)`)

# round Avg_Marriage_Rate till 2 decimal places
data5$Avg_Marriage_Rate <- round(data5$Avg_Marriage_Rate,2)

# display data using datatable
datatable(data5,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	Division	Year	Avg_Marriage_Rate
1	East North Central	1990	8.7
2	East North Central	1995	7.56
3	East North Central	1999	7.28
4	East North Central	2000	7.2
5	East North Central	2001	7.1
6	East North Central	2002	6.86
7	East North Central	2003	6.56
8	East North Central	2004	6.6
9	East North Central	2005	6.3
10	East North Central	2006	6.28

Showing 1 to 10 of 180 entries Previous 1 2 3 4 5 … 18 Next

```
ggplot(data5, aes(x = reorder(Year, desc(Year)), y = Avg_Marriage_Rate)) +
  geom_bar(stat = "identity",fill = "steelblue") + facet_grid(~Division) + coord_flip() +
  xlab("Year") + ylab("Average Marriage Rate")+ggtitle("Yealy Average Marriage Rate over Division") +
  theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "white", color = 
  geom_text(aes( y = Avg_Marriage_Rate,label=Avg_Marriage_Rate), hjust = -0.20, color="black", size=3.5
```

4.2 Yealy Average Marriage Rate over Region

```
# group by data by Division and Year
data6 <- data.frame(data) %>% group_by(Region, Year) %>% summarise(mean(Marriage_Rate)) %>% filter(!is.

# rename Avg Marriage Rate column
data6 <- data6 %>% rename( Avg_Marriage_Rate = `mean(Marriage_Rate)`)

# round Avg_Marriage_Rate till 2 decimal places
data6$Avg_Marriage_Rate <- round(data6$Avg_Marriage_Rate,2)

# display data using datatable
```

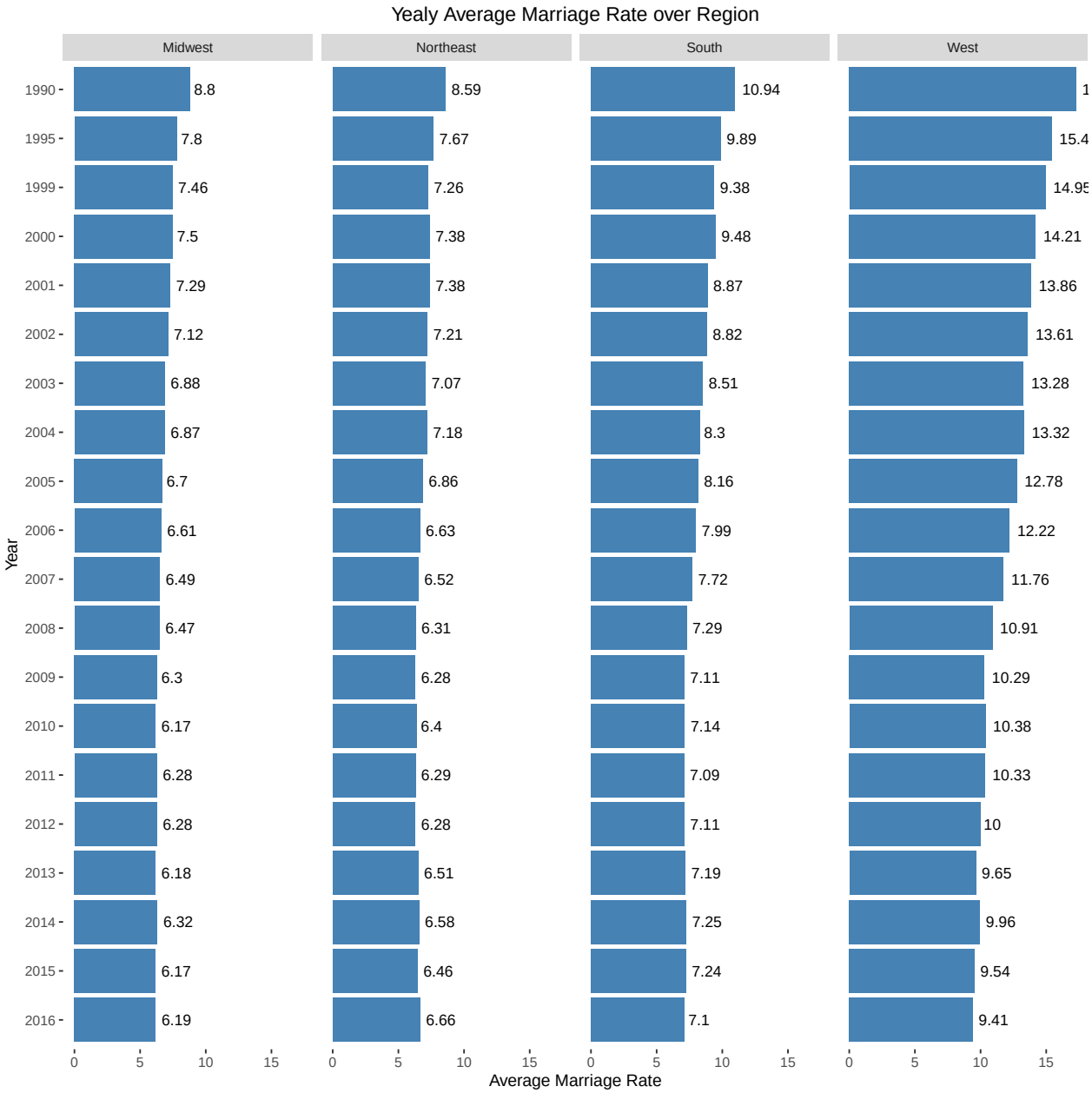
```
datatable(data6,options = list(scrollX = TRUE, paging=TRUE,fixedHeader=TRUE))
```

Show 10 entries Search:

	Region	Year	Avg_Marriage_Rate
1	Midwest	1990	8.8
2	Midwest	1995	7.8
3	Midwest	1999	7.46
4	Midwest	2000	7.5
5	Midwest	2001	7.29
6	Midwest	2002	7.12
7	Midwest	2003	6.88
8	Midwest	2004	6.87
9	Midwest	2005	6.7
10	Midwest	2006	6.61

Showing 1 to 10 of 80 entries Previous 1 2 3 4 5 … 8 Next

```
ggplot(data6, aes(x = reorder(Year, desc(Year)), y = Avg_Marriage_Rate)) +
  geom_bar(stat = "identity",fill = "steelblue") + facet_grid(~Region) + coord_flip() +
  xlab("Year") + ylab("Average Marriage Rate")+ggtitle("Yealy Average Marriage Rate over Region") +
  theme(plot.title = element_text(hjust = 0.5),panel.background = element_rect(fill = "white", color = 
  geom_text(aes( y = Avg_Marriage_Rate,label=Avg_Marriage_Rate), hjust = -0.20, color="black", size=3.5
```



5. Conclusion

The plot 4.1 and plot 4.2 shows the Average Marriage Rate is decrasing from year 1990 to 2016.

- Yealy Average Marriage Rate over Division: The Average Marriage Rate in `Mountain` division has decreased from 21.5 to 10.01 which is a decrease of 11.5%. The Average Marriage Rate in `Middle Atlantic` division has decreased from 7.7 to 6.29 which is a decrease of 1.4%. So, the Average Marriage Rate in `Middle Atlantic` division decreased less as compared to `Mountain` division.
- Yealy Average Marriage Rate over Region: The Average Marriage Rate in `West`region has decreased from 17.3 to 9.41 which is a decrease of 7.9%. The Average Marriage Rate in `Midwest` region has decreased from 8.8 to 6.19 which is a decrease of 2.6%. So, the Average Marriage Rate in `Midwest` region decreased less as comparaed to `West` region.