```r
# libraries
library(dplyr)
library(ggplot2)
library(scales)
library(GGally)
library(ggthemes)
library(RColorBrewer)
library(ggmap)
library(maps)
library(mapdata)



# data
ames <- read.csv("ames_student (1).csv")


# dataset shape
dim(ames)
## [1] 2053   81
```

```r
glimpse(ames)
## Rows: 2,053
## Columns: 81
## $ MS_SubClass        <chr> "One_Story_1946_and_Newer_All_Styles", "One_Story_1~
## $ MS_Zoning          <chr> "Residential_Low_Density", "Residential_High_Densit~
## $ Lot_Frontage       <int> 141, 80, 81, 93, 74, 78, 43, 39, 0, 85, 0, 47, 152,~
## $ Lot_Area           <int> 31770, 11622, 14267, 11160, 13830, 9978, 5005, 5389~
## $ Street             <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pa~
## $ Alley              <chr> "No_Alley_Access", "No_Alley_Access", "No_Alley_Acc~
## $ Lot_Shape          <chr> "Slightly_Irregular", "Regular", "Slightly_Irregula~
## $ Land_Contour       <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "HLS", "L~
## $ Utilities          <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "~
## $ Lot_Config         <chr> "Corner", "Inside", "Corner", "Corner", "Inside", "~
```

```
## $ Land_Slope         <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "G~
## $ Neighborhood       <chr> "North_Ames", "North_Ames", "North_Ames", "North_Am~
## $ Condition_1        <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "N~
## $ Condition_2        <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "No~
## $ Bldg_Type          <chr> "OneFam", "OneFam", "OneFam", "OneFam", "OneFam", "~
## $ House_Style        <chr> "One_Story", "One_Story", "One_Story", "One_Story",~
## $ Overall_Qual       <chr> "Above_Average", "Average", "Above_Average", "Good"~
## $ Overall_Cond       <chr> "Average", "Above_Average", "Above_Average", "Avera~
## $ Year_Built         <int> 1960, 1961, 1958, 1968, 1997, 1998, 1992, 1995, 199~
## $ Year_Remod_Add     <int> 1960, 1961, 1958, 1968, 1998, 1998, 1992, 1996, 200~
## $ Roof_Style         <chr> "Hip", "Gable", "Hip", "Hip", "Gable", "Gable", "Ga~
## $ Roof_Matl          <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompSh~
## $ Exterior_1st       <chr> "BrkFace", "VinylSd", "Wd Sdng", "BrkFace", "VinylS~
## $ Exterior_2nd       <chr> "Plywood", "VinylSd", "Wd Sdng", "BrkFace", "VinylS~
## $ Mas_Vnr_Type       <chr> "Stone", "None", "BrkFace", "None", "None", "BrkFac~
## $ Mas_Vnr_Area       <int> 112, 0, 108, 0, 0, 20, 0, 0, 0, 0, 0, 603, 0, 350, ~
## $ Exter_Qual         <chr> "Typical", "Typical", "Typical", "Good", "Typical",~
## $ Exter_Cond         <chr> "Typical", "Typical", "Typical", "Typical", "Typica~
## $ Foundation         <chr> "CBlock", "CBlock", "CBlock", "CBlock", "PConc", "P~
## $ Bsmt_Qual          <chr> "Typical", "Typical", "Typical", "Typical", "Good",~
## $ Bsmt_Cond          <chr> "Good", "Typical", "Typical", "Typical", "Typical",~
## $ Bsmt_Exposure      <chr> "Gd", "No", "No", "No", "No", "No", "No", "No", "No~
## $ BsmtFin_Type_1     <chr> "BLQ", "Rec", "ALQ", "ALQ", "GLQ", "GLQ", "ALQ", "G~
## $ BsmtFin_SF_1       <int> 2, 6, 1, 1, 3, 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, ~
## $ BsmtFin_Type_2     <chr> "Unf", "LwQ", "Unf", "Unf", "Unf", "Unf", "Unf", "U~
## $ BsmtFin_SF_2       <int> 0, 144, 0, 0, 0, 0, 0, 0, 0, 0, 1120, 0, 0, 0, 0, 0~
## $ Bsmt_Unf_SF        <int> 441, 270, 406, 1045, 137, 324, 1017, 415, 233, 663,~
## $ Total_Bsmt_SF      <int> 1080, 882, 1329, 2110, 928, 926, 1280, 1595, 1168, ~
## $ Heating            <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", "Ga~
## $ Heating_QC         <chr> "Fair", "Typical", "Typical", "Excellent", "Good", ~
## $ Central_Air        <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "~
## $ Electrical         <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr~
## $ First_Flr_SF       <int> 1656, 896, 1329, 2110, 928, 926, 1280, 1616, 1187, ~
## $ Second_Flr_SF      <int> 0, 0, 0, 0, 701, 678, 0, 0, 0, 0, 0, 1589, 672, 0, ~
```

```
## $ Low_Qual_Fin_SF    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Gr_Liv_Area        <int> 1656, 896, 1329, 2110, 1629, 1604, 1280, 1616, 1187~
## $ Bsmt_Full_Bath     <int> 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, ~
## $ Bsmt_Half_Bath     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Full_Bath          <int> 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1, 2, 2, 1, ~
## $ Half_Bath          <int> 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, ~
## $ Bedroom_AbvGr      <int> 3, 2, 3, 3, 3, 3, 2, 2, 3, 2, 1, 4, 4, 1, 3, 3, 2, ~
## $ Kitchen_AbvGr      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Kitchen_Qual       <chr> "Typical", "Typical", "Good", "Excellent", "Typical~
## $ TotRms_AbvGrd      <int> 7, 5, 6, 8, 6, 7, 5, 5, 6, 5, 4, 12, 8, 8, 7, 7, 5,~
## $ Functional         <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "T~
## $ Fireplaces         <int> 2, 0, 0, 2, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, ~
## $ Fireplace_Qu       <chr> "Good", "No_Fireplace", "No_Fireplace", "Typical", ~
## $ Garage_Type        <chr> "Attchd", "Attchd", "Attchd", "Attchd", "Attchd", "~
## $ Garage_Finish      <chr> "Fin", "Unf", "Unf", "Fin", "Fin", "Fin", "RFn", "R~
## $ Garage_Cars        <int> 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, ~
## $ Garage_Area        <int> 528, 730, 312, 522, 482, 470, 506, 608, 420, 506, 5~
## $ Garage_Qual        <chr> "Typical", "Typical", "Typical", "Typical", "Typica~
## $ Garage_Cond        <chr> "Typical", "Typical", "Typical", "Typical", "Typica~
## $ Paved_Drive        <chr> "Partial_Pavement", "Paved", "Paved", "Paved", "Pav~
## $ Wood_Deck_SF       <int> 210, 140, 393, 0, 212, 360, 0, 237, 483, 192, 0, 50~
## $ Open_Porch_SF      <int> 62, 0, 36, 0, 34, 36, 82, 152, 21, 0, 54, 36, 12, 0~
## $ Enclosed_Porch     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Three_season_porch <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Screen_Porch       <int> 0, 120, 0, 0, 0, 0, 144, 0, 0, 0, 140, 210, 0, 0, 0~
## $ Pool_Area          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Pool_QC            <chr> "No_Pool", "No_Pool", "No_Pool", "No_Pool", "No_Poo~
## $ Fence              <chr> "No_Fence", "Minimum_Privacy", "No_Fence", "No_Fenc~
## $ Misc_Feature       <chr> "None", "None", "Gar2", "None", "None", "None", "No~
## $ Misc_Val           <int> 0, 0, 12500, 0, 0, 0, 0, 0, 500, 0, 0, 0, 0, 0, 0, ~
## $ Mo_Sold            <int> 5, 6, 6, 4, 3, 6, 1, 3, 3, 2, 6, 6, 6, 6, 1, 1, 3, ~
## $ Year_Sold          <int> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 201~
## $ Sale_Type          <chr> "WD ", "WD ", "WD ", "WD ", "WD ", "WD ", "WD ", "W~
## $ Sale_Condition     <chr> "Normal", "Normal", "Normal", "Normal", "Normal", "~
```

```
## $ Longitude          <dbl> -93.61975, -93.61976, -93.61939, -93.61732, -93.638~
## $ Latitude           <dbl> 42.05403, 42.05301, 42.05266, 42.05125, 42.06090, 4~
## $ Above_Median       <chr> "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Ye~
num_vars<-colnames(ames[sapply(ames, is.numeric) == TRUE])
cat_vars<-colnames(ames[sapply(ames, is.character) == TRUE])

stmt <- paste("There are", length(num_vars), "numerical features and ", length(cat_vars), "categorical features" )
print(stmt, na.print = NULL)
## [1] "There are 34 numerical features and  47 categorical features"

# Histogram of living area
ggplot(ames, aes(x = Gr_Liv_Area)) +
 geom_histogram(color = "black", fill = "orange2", bins = 50) +
 scale_x_continuous(labels = comma) +
 labs(title = "Distribution of house sizes", x = "Living area (sqft)", y = "Frequency") +
 theme_minimal()

ames %>%
 ggplot(aes(x=Above_Median, fill=Above_Median))+
 geom_text(aes(label = ..count..), stat = "count", vjust = 0.00000005, colour = "black")+
 geom_bar(width=0.5)+ ggtitle("Barplot of Above Median")

ames %>%
 ggplot(aes(x=Year_Sold, fill=Year_Sold))+
 geom_text(aes(label = ..count..), stat = "count", vjust = 0.00000005, colour = "black")+
 geom_bar(width=0.5, fill="green1")+ ggtitle("Number of Houses sold per Year")

ames %>%
 ggplot(aes(x=as.factor(Mo_Sold)))+
 geom_text(aes(label = ..count..), stat = "count", vjust = 0.00000005, colour = "black")+
 # facet_grid(. ~ Above_Median)+
 geom_bar(width=0.5, fill="cyan")+ labs(x="Month", fill="blue")+ ggtitle("Number of Houses sold per Month")
```

```
ames %>%
 ggplot(aes(x=as.factor(Overall_Qual), fill=Overall_Qual))+
 geom_text(aes(label = ..count..), stat = "count", vjust = 0.00000005, colour = "black")+
 # facet_grid(. ~ Above_Median)+
 geom_bar(width=0.5)+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
labs(x="Quality rating", colour="blue")+ ggtitle("Number of Houses sold by House quality rating")



ames %>%
 ggplot(aes(x=as.factor(Overall_Cond), fill=Overall_Cond))+
 geom_text(aes(label = ..count..), stat = "count", vjust = 0.0000024, colour = "black")+
 # facet_grid(. ~ Above_Median)+
 geom_bar(width=0.5)+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
 labs(x="Condition rating")+ ggtitle("Number of Houses sold by House condition rating")
yr.blt <- data.frame(table(ames$Year_Built))
colnames(yr.blt) <- c("Year_Built", "Count")
yr.blt %>% arrange(desc(Count)) %>%
 ggplot(aes(x=Year_Built, y=Count))+
 # geom_text(aes(label = ..count..), stat = "count", vjust = 0.00000005, colour = "black")+
 # facet_grid(. ~ Above_Median)+
 geom_col(fill="darkblue")+
 theme_economist() +
 scale_color_economist()+
 theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=0.5, size=5))+
 labs(x="Year", colour="blue")+ ggtitle("Year of Construction of the house sold")
# compute the bounding box
bc_bbox <- make_bbox(lat = Latitude, lon = Longitude, data = ames)


# grab the maps from google
bc_big <- get_map(location = bc_bbox, source = "google", maptype = "terrain")


# plot the points and color them by sector
cnt_by_nei <-  ames %>% group_by(Neighborhood) %>%
  summarise(Latitude=mean(Latitude), Longitude=mean(Longitude), Count=n()*1000)
```

```
attach(cnt_by_nei)
ggmap(bc_big) +
 geom_point(data = cnt_by_nei, mapping = aes(x = Longitude, y = Latitude,
                                             size=Count*10000^2,color = Neighborhood))+
 theme_classic() +ggtitle("House sold count by Neighborhood")
```