Insurance

## R Markdown

The read.csv is used to import the data, and file.choose helps one bypass the troubles of setting working directory.

Let’s dig into the insurance data set.

insurance<-read.csv(file.choose(),header=T)

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(extrafont)

## Registering fonts with R

Sometimes I pick and choose between head and View. Nevertheless, I use them to have a quick preview of the data.

View(insurance)  
  
head(insurance, 10)

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622  
## 7 46 female 33.440 1 no southeast 8240.590  
## 8 37 female 27.740 3 no northwest 7281.506  
## 9 37 male 29.830 2 no northeast 6406.411  
## 10 60 female 25.840 0 no northwest 28923.137

Let’s see how R imported the data. A close look at data types

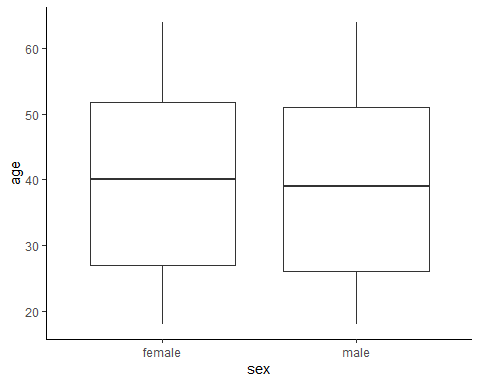
str(insurance)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : chr "female" "male" "male" "male" ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : chr "yes" "no" "no" "no" ...  
## $ region : chr "southwest" "southeast" "southeast" "northwest" ...  
## $ charges : num 16885 1726 4449 21984 3867 ...

Now some data Viz to explore what our data looks like.

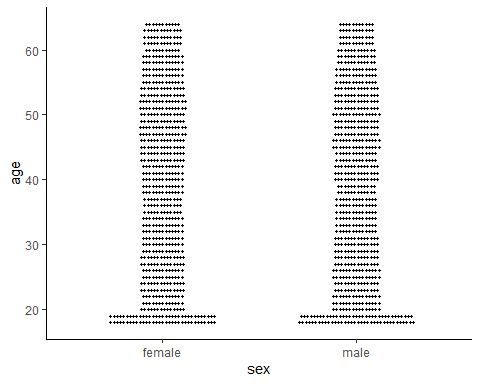
First, I would like to see the spread of age data by sex. Do we have older males or females.

ggplot(insurance)+  
 geom\_boxplot(aes(sex, age))+theme\_classic()



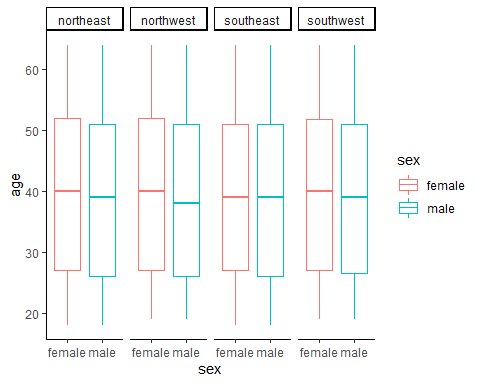
From the box plot the data the age data looks spread evenly between male and female, I will take a closer look using the dot-plot geom

ggplot(insurance)+  
 geom\_dotplot(aes(sex, age), binaxis="y", dotsize=1, binwidth=0.5, stackdir="center")+theme\_classic()

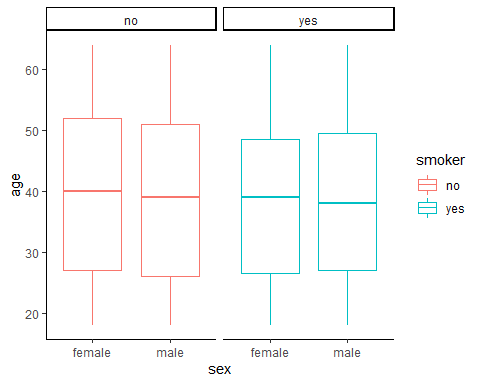
 The dotplot shoes us something interesting in the ~20 age group. Younger study population?

Time to visualize the age distribution by region and smoking

ggplot(insurance)+  
 geom\_boxplot(aes(sex, age, color=sex))+theme\_classic() +facet\_grid(~region)



ggplot(insurance)+  
 geom\_boxplot(aes(sex, age, color=smoker))+theme\_classic() +facet\_grid(~smoker)



Some summary analysis to view the charges by region, grouping the data by sex.

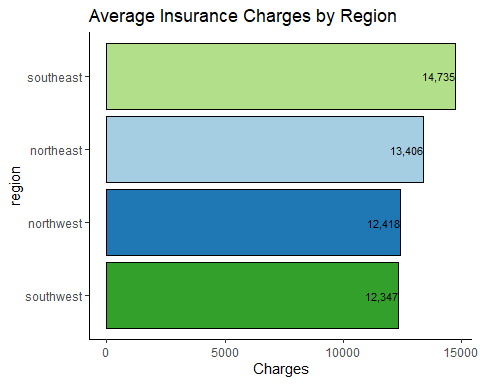
First, I will create a summary data to view the average charges by region

charges\_by\_region <- insurance %>%   
 group\_by(region) %>%  
 summarise(Avg\_region\_charges=mean(charges)) %>% ungroup  
  
charges\_by\_region

## # A tibble: 4 x 2  
## region Avg\_region\_charges  
## <chr> <dbl>  
## 1 northeast 13406.  
## 2 northwest 12418.  
## 3 southeast 14735.  
## 4 southwest 12347.

Time to visualize

charges\_by\_region %>%   
 ggplot(aes(reorder(region,Avg\_region\_charges),Avg\_region\_charges,fill=region))+  
 geom\_col(show.legend = FALSE,color="black")+  
geom\_text(aes(label=comma(Avg\_region\_charges)),size=3,hjust=1,color="black")+  
 scale\_fill\_brewer(palette = "Paired")+  
 coord\_flip()+  
 theme\_classic()+  
 labs(title = "Average Insurance Charges by Region",x="region",y= "Charges")



Next We visualize Charges by sex

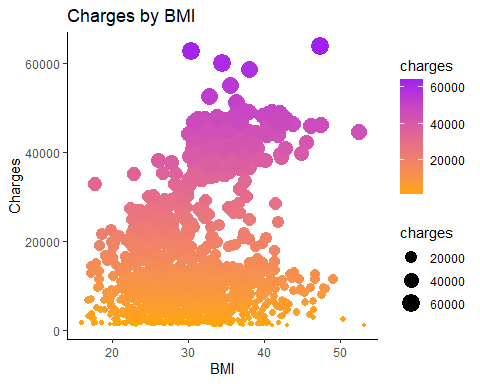
charges\_by\_sex <- insurance %>%   
 group\_by(sex) %>%  
 summarise(Avg\_region\_charges\_sex=mean(charges)) %>% ungroup  
  
charges\_by\_sex

## # A tibble: 2 x 2  
## sex Avg\_region\_charges\_sex  
## <chr> <dbl>  
## 1 female 12570.  
## 2 male 13957.

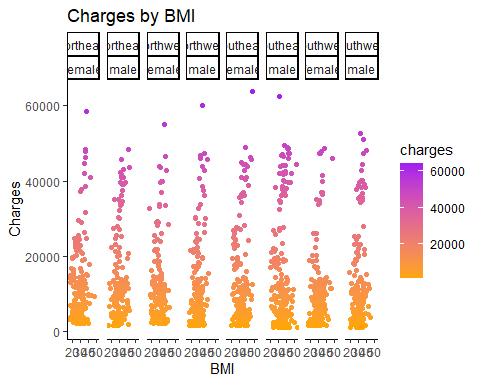
Males are charged more? Does our data have any answer to this? Do we have more males in the southeast? Does smoking status or BMI affect charges?

Lets get on some visualizations to answer these questions.

ggplot(insurance)+  
 geom\_point(aes(bmi,charges, color=charges, size=charges))+  
 scale\_color\_gradient(low="orange", high="purple")+  
 theme\_classic()+  
 labs(title = "Charges by BMI",x="BMI",y= "Charges")

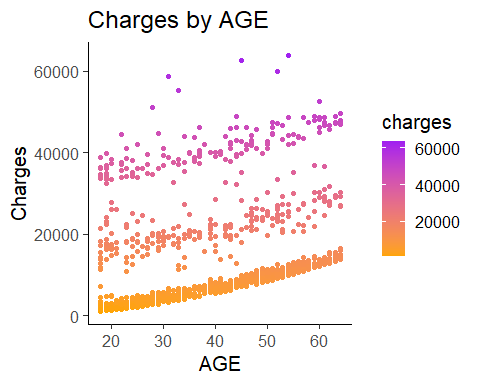
 Lets View this data by Region and Sex

ggplot(insurance)+  
 geom\_point(aes(bmi,charges, color=charges))+  
 scale\_color\_gradient(low="orange", high="purple")+  
 theme\_classic()+  
 labs(title = "Charges by BMI",x="BMI",y= "Charges") + facet\_grid(.~region+sex)



First, we see that most of the charges are below 20000; also, there is no defined relationship between charges and BMI. Can age be the answer?

ggplot(insurance)+  
 geom\_point(aes(age,charges, color=charges))+  
 scale\_color\_gradient(low="orange", high="purple")+  
 theme\_classic(base\_size=15)+  
 labs(title = "Charges by AGE",x="AGE",y= "Charges")



lmBMI = lm(charges~bmi, data = insurance)  
summary(lmBMI)

##   
## Call:  
## lm(formula = charges ~ bmi, data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20956 -8118 -3757 4722 49442   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1192.94 1664.80 0.717 0.474   
## bmi 393.87 53.25 7.397 2.46e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11870 on 1336 degrees of freedom  
## Multiple R-squared: 0.03934, Adjusted R-squared: 0.03862   
## F-statistic: 54.71 on 1 and 1336 DF, p-value: 2.459e-13

The BMI is a good addition to the model P = 2.46e-13 \*\*\*, However, the R square shows that this model is a poor fit for the data.

Now lets try charges by age

lmage = lm(charges~age, data = insurance)  
summary(lmage)

##   
## Call:  
## lm(formula = charges ~ age, data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8059 -6671 -5939 5440 47829   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3165.9 937.1 3.378 0.000751 \*\*\*  
## age 257.7 22.5 11.453 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11560 on 1336 degrees of freedom  
## Multiple R-squared: 0.08941, Adjusted R-squared: 0.08872   
## F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

Similarly, age is a good predictor for the model as the P-value is < 2e-16. However, this model explains only 9 percent of the variation in the model.

Human studies typically have low R2 because of unpredictability, but will 4 percent or 9 percent explained variation pass? I don’t think so.

I’ll try a multiple regression before I come to initial conclusions.

lmagebmi = lm(charges~age + bmi, data = insurance)  
summary(lmagebmi)

##   
## Call:  
## lm(formula = charges ~ age + bmi, data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14457 -7045 -5136 7211 48022   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6424.80 1744.09 -3.684 0.000239 \*\*\*  
## age 241.93 22.30 10.850 < 2e-16 \*\*\*  
## bmi 332.97 51.37 6.481 1.28e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11390 on 1335 degrees of freedom  
## Multiple R-squared: 0.1172, Adjusted R-squared: 0.1159   
## F-statistic: 88.6 on 2 and 1335 DF, p-value: < 2.2e-16

VIola! both age and bmi are good predictors for charges, however their R2 is only 12 percent. Is this model useable? Yes!First this data set might be a subset of a large data set and the sample size has been reduced, also, this might be a streamlined data set, where other variables that will bolster the model have been removed.

I added children to the regression and the R2 increased to 13 percent.

What is your take on this interpretation.

Final phase of working with this data, I want to answer the question ” Does region, smoker or sex” have any effect on charges

I first ensure I have my factors in place and check the data types after chnaging them.

insurance$sex=as.factor(insurance$sex)  
  
insurance$smoker=as.factor(insurance$smoker)  
  
insurance$region=as.factor(insurance$region)  
  
str(insurance)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ charges : num 16885 1726 4449 21984 3867 ...

insaov=aov(charges~(smoker\*sex)+region,data=insurance)  
anova(insaov)

## Analysis of Variance Table  
##   
## Response: charges  
## Df Sum Sq Mean Sq F value Pr(>F)   
## smoker 1 1.2152e+11 1.2152e+11 2187.0555 < 2.2e-16 \*\*\*  
## sex 1 1.4213e+06 1.4213e+06 0.0256 0.872954   
## region 3 1.0780e+08 3.5933e+07 0.6467 0.585068   
## smoker:sex 1 4.9042e+08 4.9042e+08 8.8263 0.003023 \*\*   
## Residuals 1331 7.3955e+10 5.5563e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

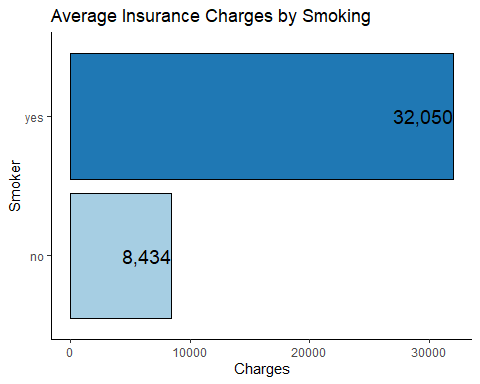
Now it all makes sense!!!!!!! Charges are statistically similar for sex and region!!! Though I blocked by region, which is a random factor. I did not expect a difference in charges by region.

Now let's do the final visualizations to see what the data looks like by smoking….

charges\_by\_smoke <- insurance %>%   
 group\_by(smoker) %>%  
 summarise(Avg\_smoke\_charges=mean(charges)) %>% ungroup  
  
charges\_by\_smoke

## # A tibble: 2 x 2  
## smoker Avg\_smoke\_charges  
## <fct> <dbl>  
## 1 no 8434.  
## 2 yes 32050.

charges\_by\_smoke %>%   
 ggplot(aes(reorder(smoker,Avg\_smoke\_charges),Avg\_smoke\_charges,fill=smoker))+  
 geom\_col(show.legend = FALSE,color="black")+  
geom\_text(aes(label=comma(Avg\_smoke\_charges)),size=5,hjust=1,color="black")+  
 scale\_fill\_brewer(palette = "Paired")+  
 coord\_flip()+  
 theme\_classic()+  
 labs(title = "Average Insurance Charges by Smoking",x="Smoker",y= "Charges")



The End on today’s episode of telling stories with datasets.