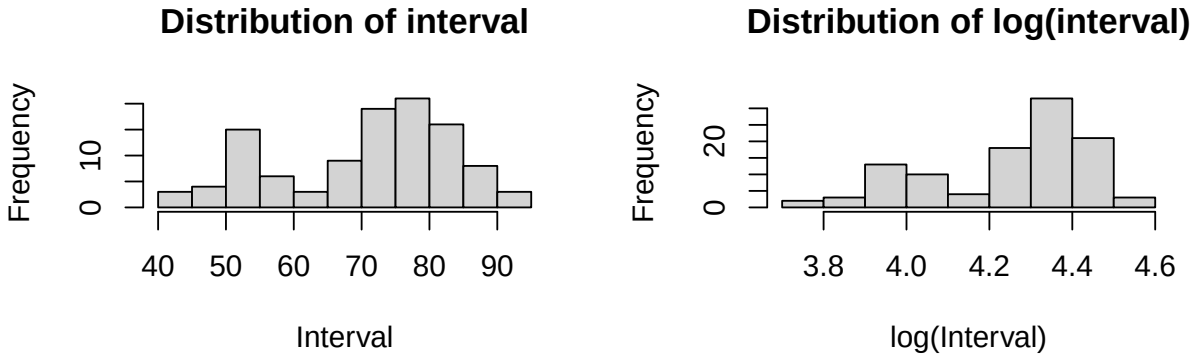
Data Analysis Assignment 2

Question1

After simple exploratory data analysis, we can tell that this dataset contains 107 rows and 4 columns(variables). Data type for four variables are X(int), Date(num), Interval(int), and Duration(num).Tables below show some statistic features of variables **Interval**, and **Duration**. Plots show the distribution of reponse variable. The distribution of interval is not very normal. After transformation to log(interval), it is still not fully normal.



Interval

Min	first_Qu	Median	Mean	third_Qu	Max
42	59	75	71	80.5	95

Duration

Min	first_Qu	Median	Mean	third_Qu	Max
1.7	2.3	3.8	3.46	4.3	4.9

Regression Model

The regression model is **Interval** $=\beta 0 + \beta 1$**Duration** $+ \epsilon$

- The results of the linear model indicate that Duration is a significant predictor of Interval at the $\alpha = 0.05$ significance level ($p < 0.001$). For every unit increase in the duration of previous one(one minute longer in duration time), the interval between eruptions increase by 10.74 minutes. When duration of precious one is zero, the interval is 33.83 minutes. (intercept has no practical value and is not meaningful enough).

	Estimate	SE	t	p-value
Intercept	33.8282	2.2618	14.9562	<.001
Duration	10.741	0.6263	17.1489	<.001

- **The 95% confidence interval for the slope is (9.50,11.98)**, which is a set of possible values of the estimated slope that we are 95% confident contains the true value for the slope parameter, meaning we are 95% confident that For every unit increase in the duration of previous one(one minute longer in duration time), the true interval between eruptions increase between 9.50 and 11.98 minutes.
- The model fit produces an R^2 value of 0.74, meaning using duration time to estimate the interval time between eruptions reduced the uncertainty in the estimate by explaining approximately 74% of the variability in the response.(74% of the variation in interval time between eruptions is explained by duration time of the previous one.)
- The regression assumptions are plausible.The trend between the response variable and predictors is roughly linear. The Q-Q plot is a little bit skewed, but the clustering of the points is still around 45 degrees, so the normality assumption is not violated. In the residuals vs fitted values plot, there seems to be a pattern, but I think the pattern is possibly due to a lack of data in certain fitted values. So there is no obvious pattern and the nearly normal residuals assumption is not violated. The variance is also roughly constant. Each eruption can be viewed as an independent observation. Therefore, the model assumptions are reasonable.

Table 3: Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.88	3.07	10.72	0.00
Duration	10.88	0.66	16.43	0.00
Date_fac2	1.33	2.72	0.49	0.63
Date_fac3	0.78	2.70	0.29	0.77
Date_fac4	0.16	2.65	0.06	0.95
Date_fac5	0.25	2.65	0.09	0.93
Date_fac6	1.99	2.66	0.75	0.46
Date_fac7	-0.17	2.70	-0.06	0.95
Date_fac8	-0.69	2.70	-0.26	0.80

The other regression model is **Interval** $=\beta 0 + \beta 1$**Duration** $+ \beta 2$**Date_fac** $+ \epsilon$

- According to the result, there is no significant difference in mean intervals for any of the days.

K-Fold Validation

The RMSE for model excluding day is 6.83, the RMSE for model including day and duration is 7.48. Since 6.83<7.48, the model excluding day appears to have higher predictive accuracy based on the average RMSE values.

Question2

SUMMARY

These days, it's widely understood that mothers who smoke during pregnancy run the risk of exposing their babies to many health problems. By using a linear regression model to analyze a subset of the data in a study that addressed the issue of pregnancy and smoking, I discovered that there is an association between smoking and birth weight, and mothers who smoke tend to give birth to babies with lower birth weights than mothers who do not smoke.

INTRODUCTION

Based on the concern that mothers who smoke have increased rates of premature delivery and low birth weights, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA was conducted and relevant data was collected. This study is an observational study since mothers decided whether or not to smoke during pregnancy, and collected data includes mothers' socioeconomic and demographic characteristics, babies' birth weight, and so on. In this analysis, I'm going to use a subset of the original data that has no missing values and no information about the babies' father. To simply analyses, I'll compare babies whose mothers smoke to babies whose mothers have never smoked; focus on the relationship between mothers' smoking status and babies' birth weight; pay attention to whether the association between smoking and birth weight differs by mother's rate; and figure out the likely range for the difference in birth weights for smokers and non-smokers.

DATA

After simple exploratory data analysis, we can tell that this dataset contains 869 rows(observations), 12 columns(variables). For the analysis, I'm not going to use variables 'id', 'gestation', 'data', and 'inc'. For the remaining variables, 'bwt.oz', 'parity', 'mage', 'mht', and 'mpregwt' are continuous variables. I transformed variables 'mrace', 'med', and 'smoke' to make them as categorical variables 'mrace_fac', 'med_fac', and 'smoke_fac'. Since there are no missing values in this dataset, we do not need to deal with missing data.

Mean of Continuous Variables

bwt.oz_mean	parity_mean	mage_mean	mht_mean	mpregwt_mean
118.3602	1.952819	27.29459	64.06904	128.4787

SD of Continuous Variables

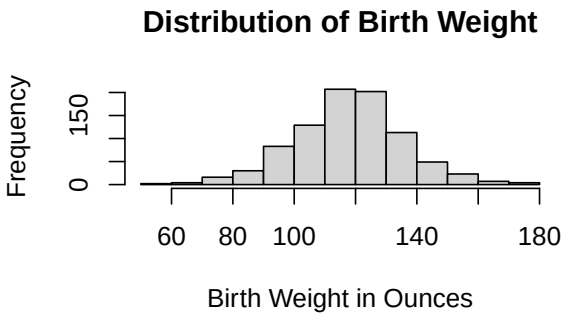
bwt.oz_sd	parity_sd	mage_sd	mht_sd	mpregwt_sd
18.05076	1.881595	5.708005	2.533612	20.77842

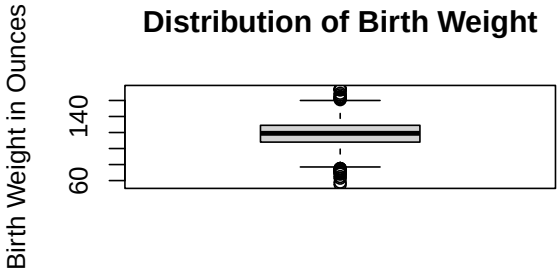
Tables of Categorical Variables

smoke_fac	n	Freq
never	466	0.5362486
smokes now	403	0.4637514

mrace_fac	n	Freq
white	626	0.7203682
mexican	25	0.0287687
black	169	0.1944764
asian	34	0.0391254
mix	15	0.0172612

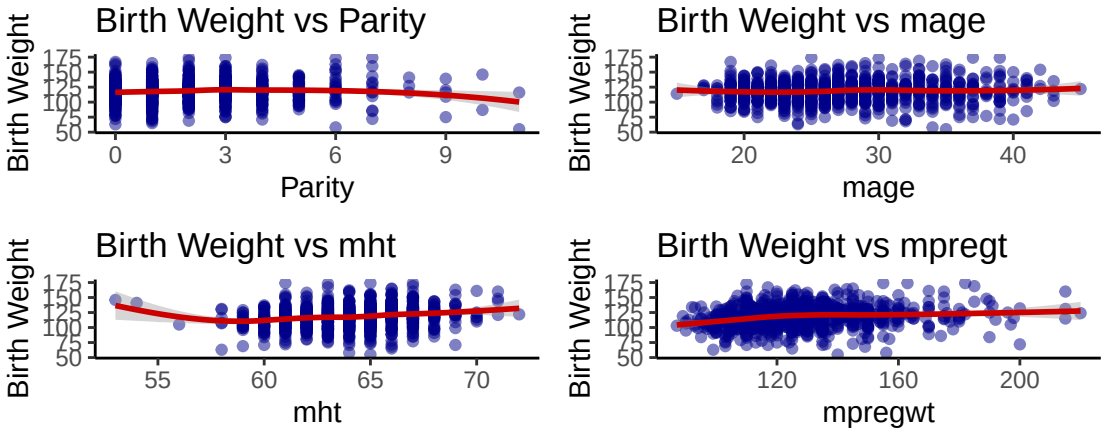
med_fac	n	Freq
< 8th	5	0.0057537
8-12th	130	0.1495972
just high school graduate	321	0.3693901
high school+trade school	47	0.0540852
high school+college	203	0.2336018
college graduate	159	0.1829689
trade school	4	0.0046030



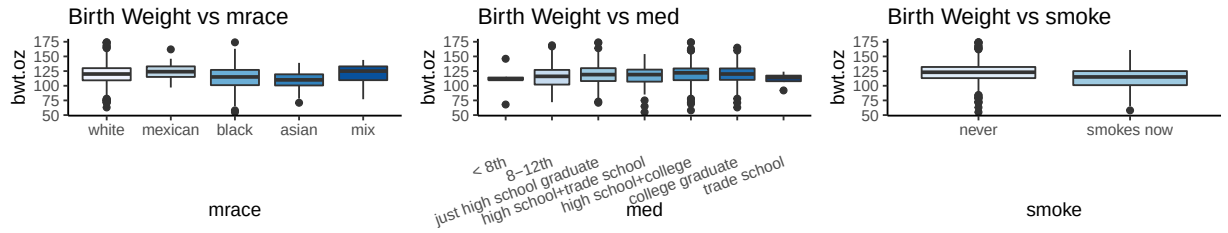


The histogram and boxplot above show that the distribution of babies' birth weight (bwz.ot) is normal.

Next, I explore the relationship between `(bwt.oz)birth weight` and each predictor. I used scatter plots for continuous/numeric predictors and boxplots for categorical predictors. The following plots show that **each trend between the response variable and predictor is roughly linear**

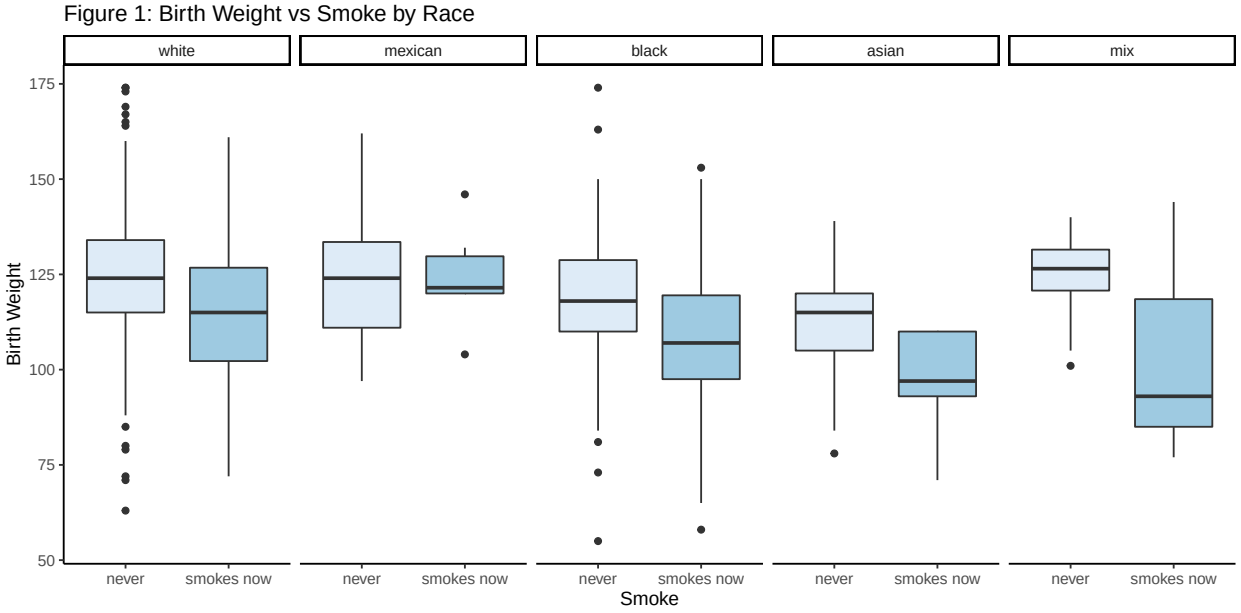


Looking at the boxplot of "Birth Weight vs smoke", we can tell that there are some difference in birthweight, that is mothers who smoke(category:smokes now) tend to give birth to babies with lower weights than mothers who do not(category:never)



After that, I tried to explore the interaction between variables, especially *whether there is any evidence that the association between smoking and birth weight differs by mothers' race.*

Give the plots below, we can tell that, **in general, in every race, mothers who smoke now tend to have babies with lower birth weights. The trend in every race is consistent. However, given those boxplots below, the difference in birth weight between mothers who smoke now and mothers who never smoke varies in every race. For example, the difference in MEXICAN is the lowest and the difference in MIX is the highest.**



I also explored interactions like "Birth Weight vs Smoke by Education", "Birth Weight vs Smoke by Mother's Age", "Birth Weight vs Smoke by Mother's Height", "Birth Weight vs Smoke by Mother's Weight", and "Birth Weight vs Smoke by Mother's Previous Pregnancy" (relevant plots are in appendix), we can tell that **all the trend in different groups are basically the same. Therefore, we can say that there is no interaction**

Model

The final regression model that I chose is $\mathbf{bwt.oz} = \beta 0 + \beta 1\mathbf{smoke\_fac(smokes\_now)} + \beta 2\mathbf{mrace\_fac(mecxican)} + \beta 3\mathbf{mrace\_fac(black)} + \beta 4\mathbf{mrace\_fac(asian)} + \beta 5\mathbf{mrace\_fac(mix)} + \beta 6\mathbf{parity} + \beta 7\mathbf{mht} + \beta 8\mathbf{mpregwt} + \epsilon$

Because we want to study the relationship between smoking and birth weights, smoke_fac is a priori that are related to the outcome of interest and must be include in the final model. Then, I used **backward**

selection using BIC to get a model $bwt.oz = \beta 0 + \beta 1 smoke\_fac(smokes\_now) + \beta 2 mrace\_fac(mecxican) + \beta 3 mrace\_fac(black) + \beta 4 mrace\_fac(asian) + \beta 5 mrace\_fac(mix) + \beta 6 mht + \beta 7 mpregwt + \epsilon$. I also used the *regsubsets function* to get the model that has the maximum adjusted R-squared: $bwt.oz = \beta 0 + \beta 1 smoke\_fac(smokes\_now) + \beta 2 mrace\_fac(mecxican) + \beta 3 mrace\_fac(black) + \beta 4 mrace\_fac(asian) + \beta 5 med\_fac(trade\ school) + \beta 6 parity + \beta 7 mht + \beta 8 mpregwt + \epsilon$ and the model that has the minimum BIC: $bwt.oz = \beta 0 + \beta 1 smoke\_fac(smokes\_now) + \beta 2 mrace\_fac(black) + \beta 3 mrace\_fac(asian) + \beta 4 mht + \beta 5 mpregwt + \epsilon$.

In addition, I did **k-fold validation. RMSE for model contains parity is 16.76, lower than 16.94, which is the RMSE for model that doesn't contain parity.**

After taking all of the above things into consideration, I think it's still reasonable to include all races in the regression model, though some of the estimated coefficients are not statistically significant. Also, I include parity(total number of previous pregnancies) into the final model since previous experience can affect how mothers behave during their current pregnancy, and the RMSE for model that contains parity is lower.

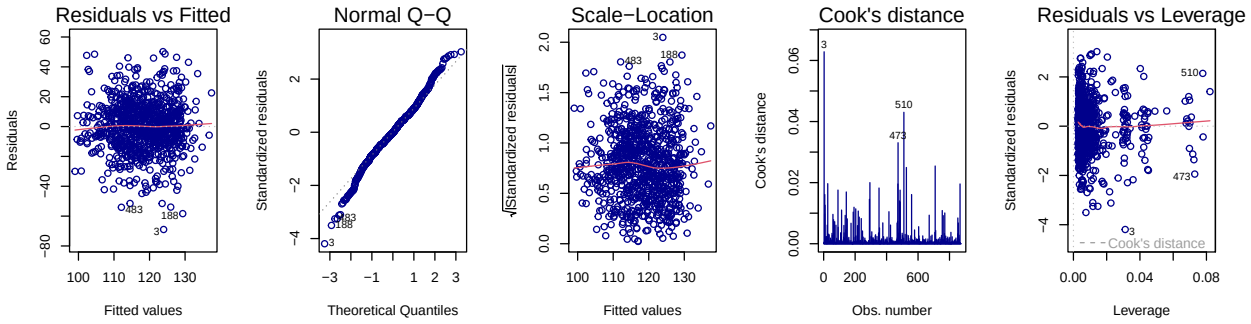
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.65	15.38	3.23	0.00
smoke_facsmokes now	-9.35	1.15	-8.12	0.00
mrace_facmexican	3.30	3.47	0.95	0.34
mrace_facblack	-8.83	1.52	-5.82	0.00
mrace_facasian	-7.94	3.04	-2.62	0.01
mrace_facmix	-1.98	4.39	-0.45	0.65
parity	0.67	0.31	2.12	0.03
mht	0.93	0.26	3.58	0.00
mpregwt	0.11	0.03	3.36	0.00

	2.5 %	97.5 %
(Intercept)	19.45	79.84
smoke_facsmokes now	-11.61	-7.09
mrace_facmexican	-3.51	10.10
mrace_facblack	-11.80	-5.85
mrace_facasian	-13.90	-1.98
mrace_facmix	-10.59	6.63
parity	0.05	1.28
mht	0.42	1.45
mpregwt	0.04	0.17

Given the regression output above, the final model is:**bwt.oz = 49.65-9.35smoke_fac(smokes_now) + 3.30mrace_fac(mecxican) -8.83mrace_fac(black)-7.94mrace_fac(asian) -1.98$mrace_fac(mix) + 0.67parity + 0.93mht + 0.11mpregwt +** ϵ

- The results of the linear model indicate that smoke_fac(smokes now), mrace_fac(black), mht, and mpregwt are significant predictors of birth weight (bwt.oz) at the $\alpha = 0.05$ significance level ($p < 0.001$).
- The intercept is 49.65, which means for mothers who never smoke, has no race, no previous pregnancy, 0 height and 0 pregnancy weight, the baby weight is estimated to be 49.65 ounces. This situation is unlikely, so the intercept is not meaningful and has no practical value.
- β_1: For mothers with the same race, same parity(previous pregnancy), same height and same weight, mothers who smoke now tend to give birth to babies that weight 9.35 ounces lower than mothers who never smoke.**The likely range for the difference in birth weights for smokers and non-smokers is (-11.61,-7.09), which means that we are 95% confident that this interval capture the true birth weight difference between smokers and never smoke mothers**
- $\beta_{2/3/4/5}$: For mothers with the same smoking status, same parity(previous pregnancy), same height and same weight, mothers who are Mexican/Black/Asian/Mix tend to give birth to babies that weight 3.30 ounces higher/8.83 ounces lower/7.94 ounces lower/1.98 ounces lower than mothers who are White. A 95% confidence interval for mrace_fac(Mexican/Black/Asian/Mix) is (-3.51,10.10)/(-11.80,-5.85)/(-13.90,-1.98)/(-10.59/6.63), meaning we are 95% confident that this interval contains the true value for the slope.
- $\beta_{6/7/8}$: For each additional number of parity(previous pregnancy)/height in inches/pre-pregnancy weight in pounds, mothers tend to give birth to babies that weight 0.67/0.93/0.11 ounces higher, holding all other variables constant. A 95% confidence interval for parity/mht/mpregwt is (0.05,1.28)/(0.42,1.45)/(0.04,0.17), meaning we are 95% confident that this interval contains the true value for the slope.
- **Adjusted R-squared is 0.14, which mean 14% of the variation in birth weight has been explained by explanatory variables. Residual standard error is 16.71 on 860 degrees of freedom, which means that the regression model predicts babies birth weight(bwt.oz) with an average error of about 16.71.**

Model Assessment



In the *DATA* part, we've already known that each trend between the response variable and predictor is roughly linear. In the residuals vs fitted values plot, though the red line is not strictly a flat line at 0, there is no obvious pattern. Also, there is independence of errors and roughly equal variance of errors. The Q-Q plot is a little bit skewed, but clustering of the points is still around the 45 degree, so normality assumption is not violated. Since **all vifs are around 1 (<10)**, we don't need to worry about multicollinearity. Besides, all points fall inside the (-3,3) standardized residuals and their cook's distance are lower than 0.5. Therefore, there aren't any (potential) outliers, leverage points or influential points.

Conclusion

Mothers who smoke tend to give birth to babies with lower weights than mothers who do not smoke. The likely range for the difference in birth weights for smokers and non-smokers is (-11.61,-7.09), and this association between smoking and birth weight does not differ by mother's race. The model can only explain 14% of the variance of the response variable(bwt.oz), which needs to be improved. The intercept now is not meaningful or realistic,methods like centering(mean-center continuous predictors) can be used to improve interpretation of the intercept.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(knitr)
library(ggplot2)
library(kableExtra)
library(lattice)
library(dplyr)
library(stargazer)
library(gt)
library(janitor)
library(flextable)
library(magrittr)
library(Hmisc)
library(caret)
library(patchwork)
library(leaps)
library(rms)
oldf <- read.csv("OldFaithful.csv")
#dim(oldf)
#head(oldf)
#summary(oldf)
#str(oldf)
par(mfrow = c(1,2))
hist(oldf$Interval,xlab="Interval",main="Distribution of interval")
hist(log(oldf$Interval),xlab="log(Interval)",main="Distribution of log(interval)")
df<-data.frame(Min=42.00,first_Qu=59.00,Median=75.00,Mean=71.00,third_Qu=80.50,Max=95.00)
df %>% regulartable%>% autofit()
df<-data.frame(Min=1.70,first_Qu=2.30,Median=3.80,Mean=3.46,third_Qu=4.30,Max=4.90)
df %>% regulartable%>% autofit()
lm <- lm(Interval ~ Duration, data = oldf)
#summary(lm)
modmat <- matrix(c(round(summary(lm)$coefficients[,1:3],4),"<.001","<.001"),nrow=2,ncol=4)
rownames(modmat) <- c("Intercept","Duration")
kable(modmat,row.names=TRUE,col.names=c("Estimate","SE","t","p-value"),format="latex",booktabs=T) %>% ka
#confint(lm, level = 0.95)
p1 <-ggplot(oldf,aes(x=Duration, y=Interval)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Interval vs Duration",x="Duration",y="Interval")
#p1
#par(mfrow = c(2,2))
#plot(lm)
oldf$Date_fac <- factor(oldf$Date,
                        levels = c("1","2","3","4","5","6","7","8"),
                        labels = c("1","2","3","4","5","6","7","8"))
lm1 <- lm(Interval ~ Duration + Date_fac, data = oldf)
knitr::kable(summary(lm1)$coefficients,digits = 2, caption = "Results")%>% kable_styling(position="cente
p2 <-ggplot(oldf,aes(x=Date_fac, y=Interval, fill=Date_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Interval vs Date_fac",x="Date_fac",y="Interval") +
```

```
  theme_classic() + theme(legend.position="none",axis.text.x = element_text(angle = 20, vjust = 0.5, hj
#p2
ctr1 <- trainControl(method = "cv", number = 10)
model <- train(Interval ~ Duration, data = oldf, method = 'lm',trContro1 = ctr1)
model1 <- train(Interval ~ Duration + Date_fac, data = oldf, method = 'lm',trContro1 = ctr1)

#knitr::kable(summary(model)$perfNames,digits = 2)%>% kable_styling(position="center",latex_options = c
#print(model)
#print(model1)
smoke <- read.csv("smoking.csv")
#smoke$smoke_fac <- factor(smoke$smoke,levels = c("0","1","2","3"),labels = c("never","smokes now","unt
smoke$smoke_fac <- factor(smoke$smoke,levels = c("0","1"),labels = c("never","smokes now"))

smoke$mrace_fac <- factor(smoke$mrace,
                          levels = c("0","1","2","3","4","5","6","7","8","9"),
                          labels = c("white","white","white","white","white","white","mexican","black",

smoke$med_fac <- factor(smoke$med,
                        levels = c("0","1","2","3","4","5","6","7"),
                        labels = c("< 8th", "8-12th","just high school graduate", "high school+trade sc

#knitr::kable(summary(smoke$smoke_fac),digits = 2)%>% kable_styling(position="center",latex_options = c
#knitr::kable(summary(smoke$mrace_fac),digits = 2)%>% kable_styling(position="center",latex_options = c
#knitr::kable(summary(smoke$med_fac),digits = 2)%>% kable_styling(position="center",latex_options = c("
df<-data.frame(bwt.oz_mean=mean(smoke$bwt.oz), parity_mean=mean(smoke$parity),mage_mean=mean(smoke$mage
df %>% regulartable%>% autofit()
df1<-data.frame(bwt.oz_sd=sd(smoke$bwt.oz), parity_sd=sd(smoke$parity),mage_sd=sd(smoke$mage),mht_sd=sd
df1 %>% regulartable%>% autofit()
t1 <- smoke%>%group_by(smoke_fac)%>%summarise(n=n())%>%mutate(Freq=n/sum(n))
t2 <- smoke%>%group_by(mrace_fac)%>%summarise(n=n())%>%mutate(Freq=n/sum(n))
t3 <- smoke%>%group_by(med_fac)%>%summarise(n=n())%>%mutate(Freq=n/sum(n))
knitr::kable(list(t1, t2,t3))
#smoke %>% janitor::tabyl(smoke_fac) %>% gt::gt()
#smoke %>% janitor::tabyl(mrace_fac) %>% gt::gt()
#smoke %>% janitor::tabyl(med_fac) %>% gt::gt()
#par(mfrow = c(1,3))
par(mfrow = c(1,2))
hist(smoke$bwt.oz,xlab="Birth Weight in Ounces",main="Distribution of Birth Weight")
boxplot(smoke$bwt.oz,ylab="Birth Weight in Ounces",main="Distribution of Birth Weight")
p1 <-ggplot(smoke,aes(x=parity, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Birth Weight vs Parity",x="Parity",y="Birth Weight")

p2 <- ggplot(smoke,aes(x=mage, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Birth Weight vs mage",x="mage",y="Birth Weight")

p3 <- ggplot(smoke,aes(x=mht, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Birth Weight vs mht",x="mht",y="Birth Weight")
```

```
p4 <- ggplot(smoke,aes(x=mpregwt, y=bwt.oz)) +
  geom_point(alpha = .5,colour="blue4") +
  geom_smooth(col="red3") + theme_classic() +
  labs(title="Birth Weight vs mpregt",x="mpregwt",y="Birth Weight")

p1+p2+p3+p4
p1 <- ggplot(smoke,aes(x=mrace_fac, y=bwt.oz, fill=mrace_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Birth Weight vs mrace",x="mrace",y="bwt.oz") +
  theme_classic() + theme(legend.position="none")

p2 <- ggplot(smoke,aes(x=med_fac, y=bwt.oz, fill=med_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Birth Weight vs med",x="med",y="bwt.oz") +
  theme_classic() + theme(legend.position="none",axis.text.x = element_text(angle = 20, vjust = 0.5, hj

p3 <- ggplot(smoke,aes(x=smoke_fac, y=bwt.oz, fill=smoke_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Birth Weight vs smoke",x="smoke",y="bwt.oz") +
  theme_classic() + theme(legend.position="none")
p1+p2+p3+plot_layout(ncol=3)
ggplot(smoke,aes(x=smoke_fac, y=bwt.oz, fill=smoke_fac)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Figure 1: Birth Weight vs Smoke by Race",x="Smoke",y="Birth Weight") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ mrace_fac,ncol = 5)
#ggplot(smoke,aes(x=smoke_fac, y=bwt.oz, fill=smoke_fac)) +
  #geom_boxplot() + #coord_flip() +
  #scale_fill_brewer(palette="Blues") +
  #labs(title="Birth Weight vs Smoke by Education",x="Smoke",y="Birth Weigght") +
  #theme_classic() + theme(legend.position="none") +
  #facet_wrap( ~ med_fac,ncol = 4)
#ggplot(smoke,aes(x=smoke_fac, y=bwt.oz)) +
  #geom_point(alpha = .5,colour="blue4") +
  #geom_smooth(method="lm",col="red3") + theme_classic() +
  #labs(title="Birth Weight vs Smoke by Mother's Age",x="Smoke",y="Birth Weight") +
  #facet_wrap( ~ mage,ncol=15)
#ggplot(smoke,aes(x=smoke_fac, y=bwt.oz)) +
  #geom_point(alpha = .5,colour="blue4") +
  #geom_smooth(method="lm",col="red3") + theme_classic() +
  #labs(title="Birth Weight vs Smoke by Mother's Height",x="Smoke",y="Birth Weight") +
  #facet_wrap( ~ mht,ncol=9)
#ggplot(smoke,aes(x=smoke_fac, y=bwt.oz)) +
  #geom_point(alpha = .5,colour="blue4") +
  #geom_smooth(method="lm",col="red3") + theme_classic() +
  #labs(title="Birth Weight vs Smoke by Mother's Weight",x="Smoke",y="Birth Weight") +
  #facet_wrap( ~ mpregwt)
#ggplot(smoke,aes(x=smoke_fac, y=bwt.oz)) +
  #geom_point(alpha = .5,colour="blue4") +
```

```
  #geom_smooth(method="lm",col="red3") + theme_classic() +
  #labs(title="Birth Weight vs Smoke by Mother's Previous Pregnancy",x="Smoke",y="Birth Weight") +
  #facet_wrap( ~ parity,ncol=6)
ctr1 <- trainControl(method = "cv", number = 10)
model <- train(bwt.oz ~ smoke_fac + mrace_fac + mht + mpregwt, data = smoke, method = 'lm',trContro1 = c
model1 <- train(bwt.oz ~ smoke_fac+mrace_fac+parity+mht+mpregwt, data = smoke, method = 'lm',trContro1 =
#print(model)
#print(model1)
fullmodel<- lm(bwt.oz ~ smoke_fac+mrace_fac+med_fac+parity+mage+mht+mpregwt, data = smoke)
#summary(fullmodel)
nullmodel<-lm(bwt.oz ~ smoke_fac, data = smoke)
#summary(nullmodel)
model_backward <- step(fullmodel,direction="backward",trace=0,k=log(869))
#model_backward$call
#summary(model_backward)
#confint(model_backward,level=0.95)
model_backward1 <- regsubsets(bwt.oz ~ smoke_fac+mrace_fac+med_fac+parity+mage+mht+mpregwt, data = smok
select_results <- summary(model_backward1)
#coef(model_backward1, which.max(select_results$adjr2))
#coef(model_backward1, which.min(select_results$bic))
model_final <- lm(bwt.oz ~ smoke_fac+mrace_fac+parity+mht+mpregwt, data = smoke)
#summary(model_final)
knitr::kable(summary(model_final)$coefficients,digits = 2)%>% kable_styling(position="center",full_widt
knitr::kable(confint(model_final,level=0.95),digits = 2)%>% kable_styling(position="center",full_width =
par(mfrow = c(1,5))
plot(model_final,which=1:5,col=c("blue4"))
#vif(model_final)
#knitr::kable(vif(model_final),digits = 2)%>% kable_styling(full_width = FALSE,latex_options = c("hold_
```