Multivariate Data Analysis - Presentation

## Multivariate Data Analysis Spring 2019 (37459-2019-SPRING-CITY)

### Assignment: Presentation

### Student Name: Anuj Kapil

### Student Id: 12678708

## Step 1: Download the dataset from Kaggle

To simplify the process, I am using my github account to host the dataset but the original dataset is available here: [Kaggle: House Sales in King County](https://www.kaggle.com/harlfoxem/housesalesprediction)

wd <- getwd()  
filep <- "https://raw.githubusercontent.com/anuj-kapil/mda/master/Data/kc\_house\_data.csv"  
filename <- "kc\_house\_data.csv"  
download.file(url=filep, destfile=filename)

## Step 2: Read the downloaded file to memory

We are using the data.table package from R to read the file in to memory

kc\_houses<-fread('kc\_house\_data.csv')

## Step 3: Exploratory Analysis

Let’s have a look at the number of features and observations in the dataset

# Rows and columns  
dim(kc\_houses)

## [1] 21597 21

Quick look at the data types and ranges of data. Below summary shows that there are no null or missing values

# No missing values  
summary(kc\_houses)

## id date price   
## Min. : 1000102 Length:21597 Min. : 78000   
## 1st Qu.:2123049175 Class :character 1st Qu.: 322000   
## Median :3904930410 Mode :character Median : 450000   
## Mean :4580474287 Mean : 540297   
## 3rd Qu.:7308900490 3rd Qu.: 645000   
## Max. :9900000190 Max. :7700000   
## bedrooms bathrooms sqft\_living sqft\_lot   
## Min. : 1.000 Min. :0.500 Min. : 370 Min. : 520   
## 1st Qu.: 3.000 1st Qu.:1.750 1st Qu.: 1430 1st Qu.: 5040   
## Median : 3.000 Median :2.250 Median : 1910 Median : 7618   
## Mean : 3.373 Mean :2.116 Mean : 2080 Mean : 15099   
## 3rd Qu.: 4.000 3rd Qu.:2.500 3rd Qu.: 2550 3rd Qu.: 10685   
## Max. :33.000 Max. :8.000 Max. :13540 Max. :1651359   
## floors waterfront view condition   
## Min. :1.000 Min. :0.000000 Min. :0.0000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.00   
## Median :1.500 Median :0.000000 Median :0.0000 Median :3.00   
## Mean :1.494 Mean :0.007547 Mean :0.2343 Mean :3.41   
## 3rd Qu.:2.000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.00   
## Max. :3.500 Max. :1.000000 Max. :4.0000 Max. :5.00   
## grade sqft\_above sqft\_basement yr\_built   
## Min. : 3.000 Min. : 370 Min. : 0.0 Min. :1900   
## 1st Qu.: 7.000 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951   
## Median : 7.000 Median :1560 Median : 0.0 Median :1975   
## Mean : 7.658 Mean :1789 Mean : 291.7 Mean :1971   
## 3rd Qu.: 8.000 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997   
## Max. :13.000 Max. :9410 Max. :4820.0 Max. :2015   
## yr\_renovated zipcode lat long   
## Min. : 0.00 Min. :98001 Min. :47.16 Min. :-122.5   
## 1st Qu.: 0.00 1st Qu.:98033 1st Qu.:47.47 1st Qu.:-122.3   
## Median : 0.00 Median :98065 Median :47.57 Median :-122.2   
## Mean : 84.46 Mean :98078 Mean :47.56 Mean :-122.2   
## 3rd Qu.: 0.00 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.:-122.1   
## Max. :2015.00 Max. :98199 Max. :47.78 Max. :-121.3   
## sqft\_living15 sqft\_lot15   
## Min. : 399 Min. : 651   
## 1st Qu.:1490 1st Qu.: 5100   
## Median :1840 Median : 7620   
## Mean :1987 Mean : 12758   
## 3rd Qu.:2360 3rd Qu.: 10083   
## Max. :6210 Max. :871200

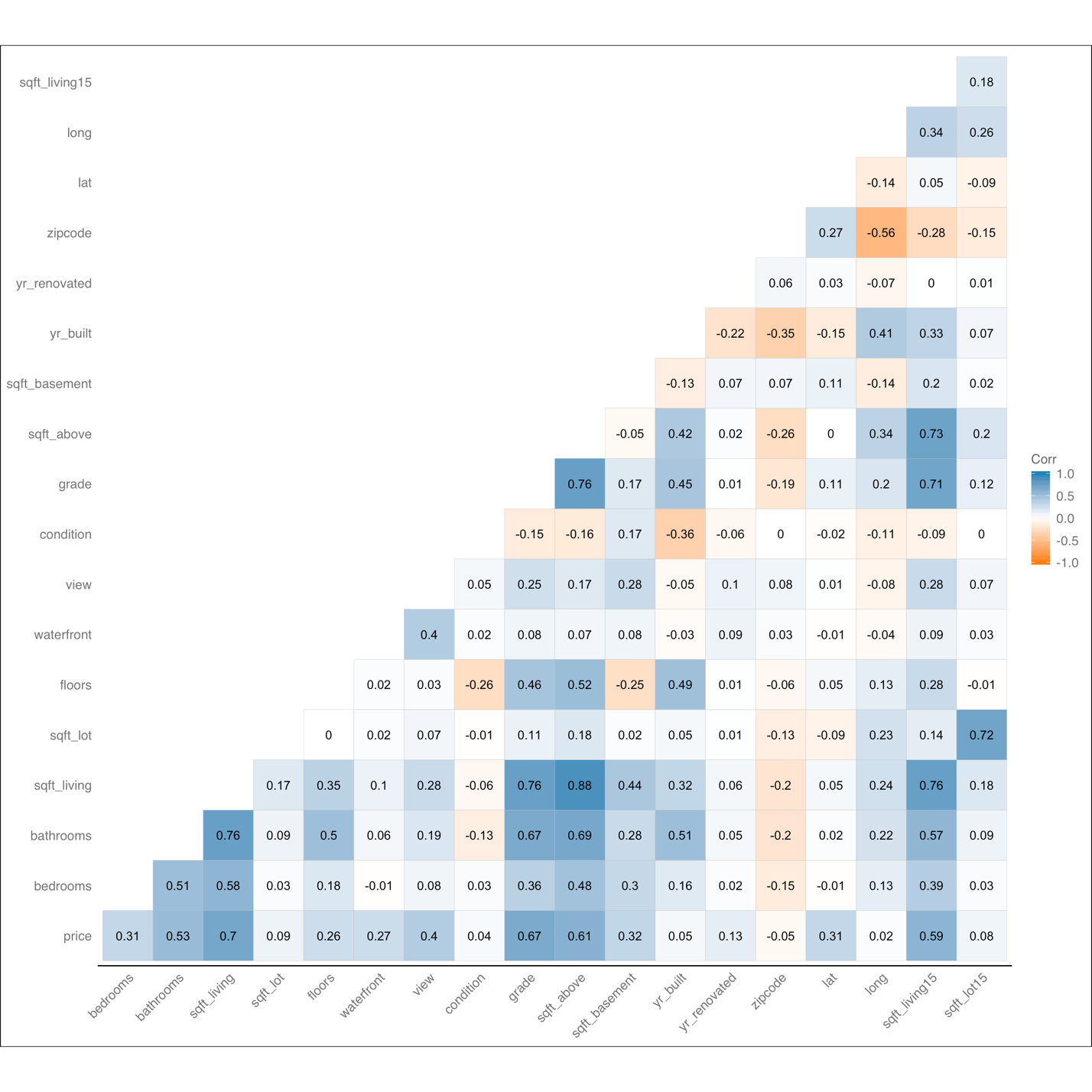
The dates are loaded as string of characters. Let’s convert them to look at the range of values. The summary now shows that the dates are in the range of May 2014 - May 2015

# Convert Dates  
kc\_houses$date<-as.IDate(kc\_houses$date, format = "%m/%d/%Y")  
summary(kc\_houses)

## id date price   
## Min. : 1000102 Min. :2014-05-02 Min. : 78000   
## 1st Qu.:2123049175 1st Qu.:2014-07-22 1st Qu.: 322000   
## Median :3904930410 Median :2014-10-16 Median : 450000   
## Mean :4580474287 Mean :2014-10-29 Mean : 540297   
## 3rd Qu.:7308900490 3rd Qu.:2015-02-17 3rd Qu.: 645000   
## Max. :9900000190 Max. :2015-05-27 Max. :7700000   
## bedrooms bathrooms sqft\_living sqft\_lot   
## Min. : 1.000 Min. :0.500 Min. : 370 Min. : 520   
## 1st Qu.: 3.000 1st Qu.:1.750 1st Qu.: 1430 1st Qu.: 5040   
## Median : 3.000 Median :2.250 Median : 1910 Median : 7618   
## Mean : 3.373 Mean :2.116 Mean : 2080 Mean : 15099   
## 3rd Qu.: 4.000 3rd Qu.:2.500 3rd Qu.: 2550 3rd Qu.: 10685   
## Max. :33.000 Max. :8.000 Max. :13540 Max. :1651359   
## floors waterfront view condition   
## Min. :1.000 Min. :0.000000 Min. :0.0000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.00   
## Median :1.500 Median :0.000000 Median :0.0000 Median :3.00   
## Mean :1.494 Mean :0.007547 Mean :0.2343 Mean :3.41   
## 3rd Qu.:2.000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.00   
## Max. :3.500 Max. :1.000000 Max. :4.0000 Max. :5.00   
## grade sqft\_above sqft\_basement yr\_built   
## Min. : 3.000 Min. : 370 Min. : 0.0 Min. :1900   
## 1st Qu.: 7.000 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951   
## Median : 7.000 Median :1560 Median : 0.0 Median :1975   
## Mean : 7.658 Mean :1789 Mean : 291.7 Mean :1971   
## 3rd Qu.: 8.000 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997   
## Max. :13.000 Max. :9410 Max. :4820.0 Max. :2015   
## yr\_renovated zipcode lat long   
## Min. : 0.00 Min. :98001 Min. :47.16 Min. :-122.5   
## 1st Qu.: 0.00 1st Qu.:98033 1st Qu.:47.47 1st Qu.:-122.3   
## Median : 0.00 Median :98065 Median :47.57 Median :-122.2   
## Mean : 84.46 Mean :98078 Mean :47.56 Mean :-122.2   
## 3rd Qu.: 0.00 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.:-122.1   
## Max. :2015.00 Max. :98199 Max. :47.78 Max. :-121.3   
## sqft\_living15 sqft\_lot15   
## Min. : 399 Min. : 651   
## 1st Qu.:1490 1st Qu.: 5100   
## Median :1840 Median : 7620   
## Mean :1987 Mean : 12758   
## 3rd Qu.:2360 3rd Qu.: 10083   
## Max. :6210 Max. :871200

## Multicollinearity Analysis

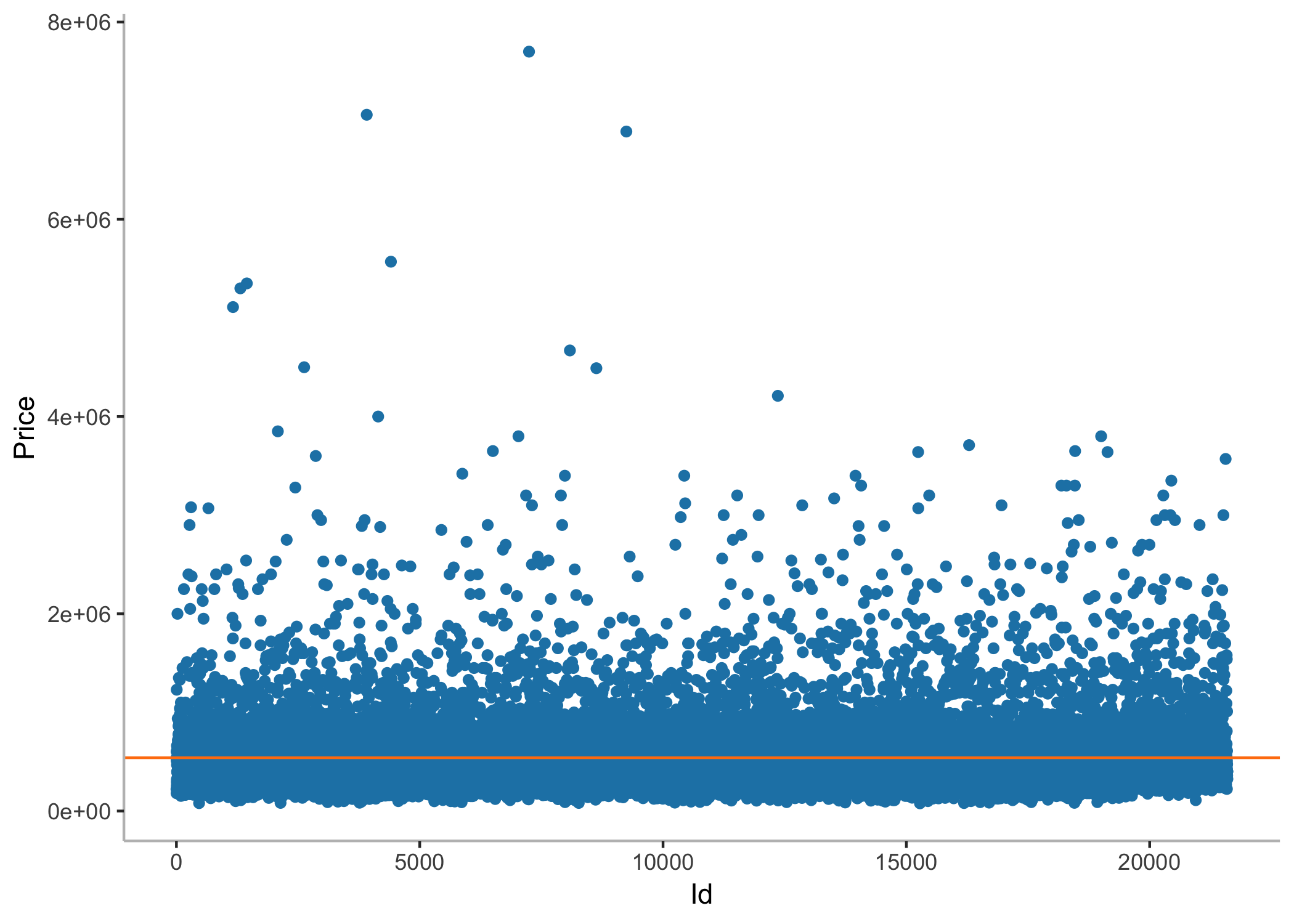
#Corr Plot  
cor\_mat <- kc\_houses[,3:21]  
corr <- cor(cor\_mat, use = "pairwise.complete.obs")  
  
ggcorrplot(corr, hc.order = FALSE, type = "lower",  
 ggtheme = ggthemes::theme\_gdocs,  
 colors = c("#ff7f0e", "white", "#1f83b4"),  
 lab = TRUE)+  
 theme(panel.grid.major=element\_blank())



## Univariate Analysis

In case of just the dependent variable and 0 independent variable, the best fit model is the mean of the dependent variable.

# Univariate model  
ggplot(kc\_houses, aes(x=seq\_along(id), y=price))+  
 geom\_point(col = "#1f83b4")+  
 geom\_hline(aes(yintercept = mean(kc\_houses$price,na.rm = T)), col = "#ff7f0e")+   
 scale\_color\_tableau() +  
 labs(x="Id", y="Price")+  
 theme(panel.background = element\_blank(), axis.line = element\_line(colour = "grey"), plot.title = element\_text(hjust = 0.5))



Calculating the Sum of Square of Errors/Residuals

kc\_houses[,err:=price-mean(price,na.rm = T)]  
kc\_houses[,err\_sq:=err^2]  
sse <- sum(kc\_houses$err\_sq)  
options(scipen = 999)  
# Total Sum of Squares = Sum of Square of Residuals  
sse

## [1] 2914582130408202

# Sample mean  
mean(kc\_houses$price,na.rm = T)

## [1] 540296.6

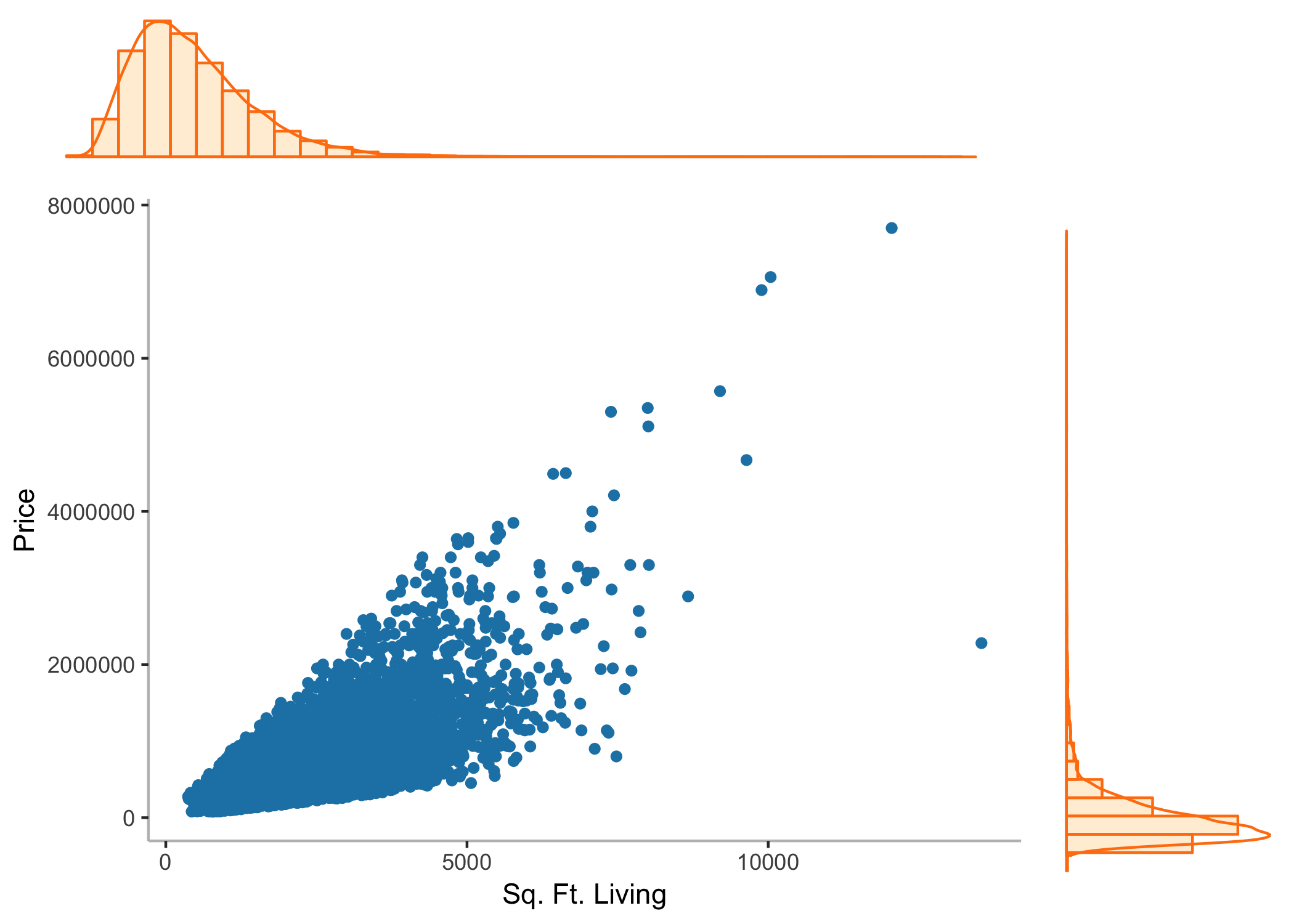
## Simple Linear Regression

Let us look at scatter plots of dependent variables and 3 chosen independent variables. Below is the function created to combine the scatter plot, histograms and density plots.

ggplot\_scatter\_hist\_combo <- function(dat, a, b, lab\_a, lab\_b){  
 # if(is.null(lab\_a))  
 # lab\_a<-a  
 # if(is.null(lab\_b))  
 # lab\_b<-b  
  
 theme\_blank<- theme(axis.line=element\_blank(),axis.text.x=element\_blank(),  
 axis.text.y=element\_blank(),axis.ticks=element\_blank(),  
 axis.title.x=element\_blank(),  
 axis.title.y=element\_blank(),legend.position="none",  
 panel.background=element\_rect(fill = "transparent"),panel.border=element\_blank(),panel.grid.major=element\_blank(),  
 panel.grid.minor=element\_blank(),plot.background=element\_rect(fill = "transparent", color = NA))  
   
 hist\_top<-ggplot(dat, aes(x=a)) +   
 geom\_histogram(aes(y =..density..),   
 col="#ff7f0e",   
 fill="#ffaa0e",   
 alpha = .2,  
 bins = 35) +   
 geom\_density(col="#ff7f0e") +   
 scale\_color\_tableau() +  
 theme\_blank  
   
 empty <- ggplot()+geom\_point(aes(1,1), colour="white")+  
 theme\_blank  
   
 scatter<-ggplot(dat, aes(x=a, y=b))+  
 geom\_point(col="#1f83b4")+  
 labs(x=lab\_a, y=lab\_b)+  
 theme(panel.background = element\_blank(), axis.line = element\_line(colour = "grey"), plot.title = element\_text(hjust = 0.5))  
   
 hist\_right <- ggplot(dat, aes(x=b)) +   
 geom\_histogram(aes(y =..density..),   
 col="#ff7f0e",   
 fill="#ffaa0e",   
 alpha = .2,  
 bins = 35) +   
 geom\_density(col="#ff7f0e") +   
 coord\_flip() +  
 theme\_blank  
   
 grid.arrange(hist\_top, empty, scatter, hist\_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))  
}

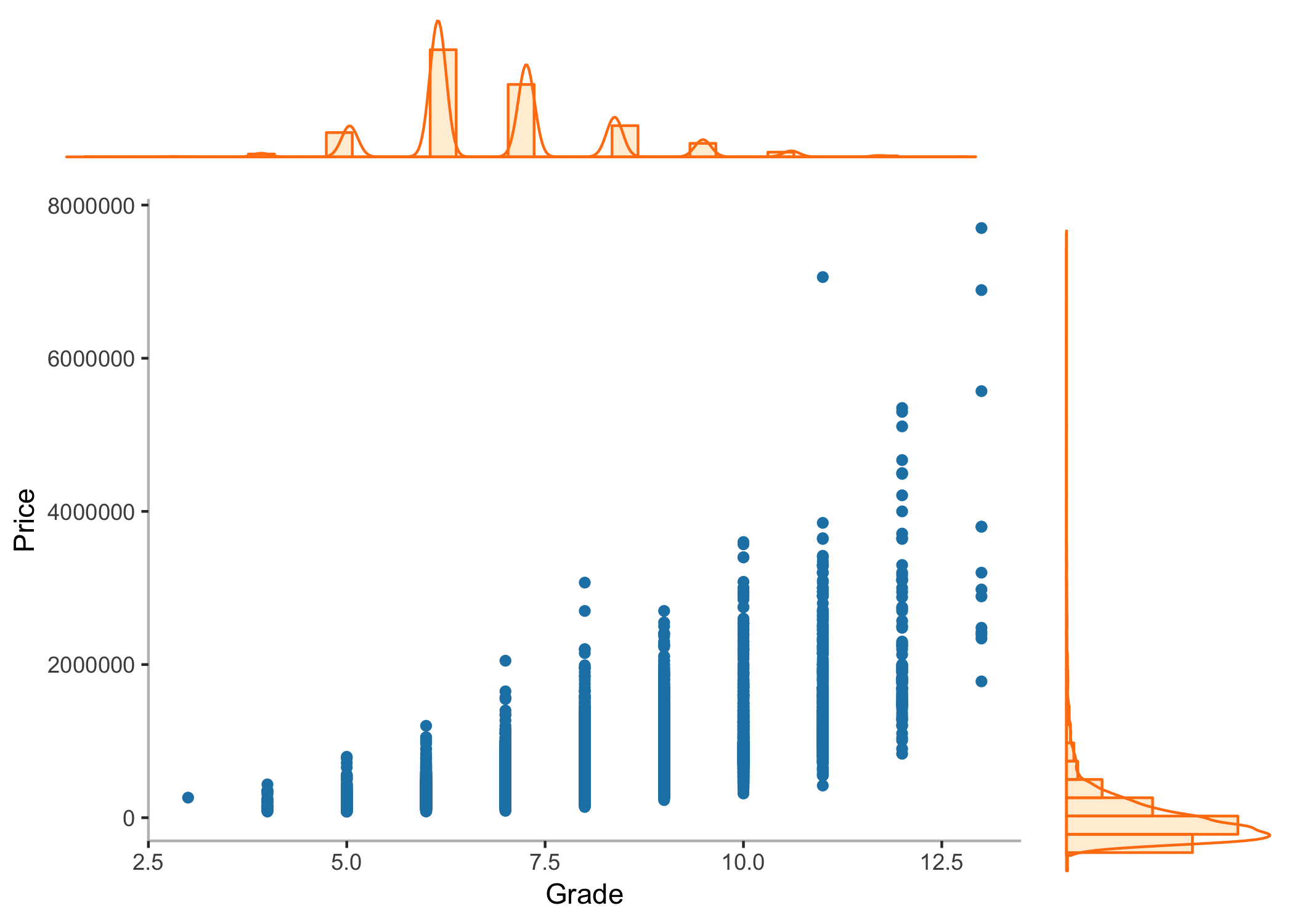
### Scatter/Histogram/Density Plots: Sq. Ft Living vs Price

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$sqft\_living, b=kc\_houses$price, lab\_a = "Sq. Ft. Living", lab\_b = "Price")



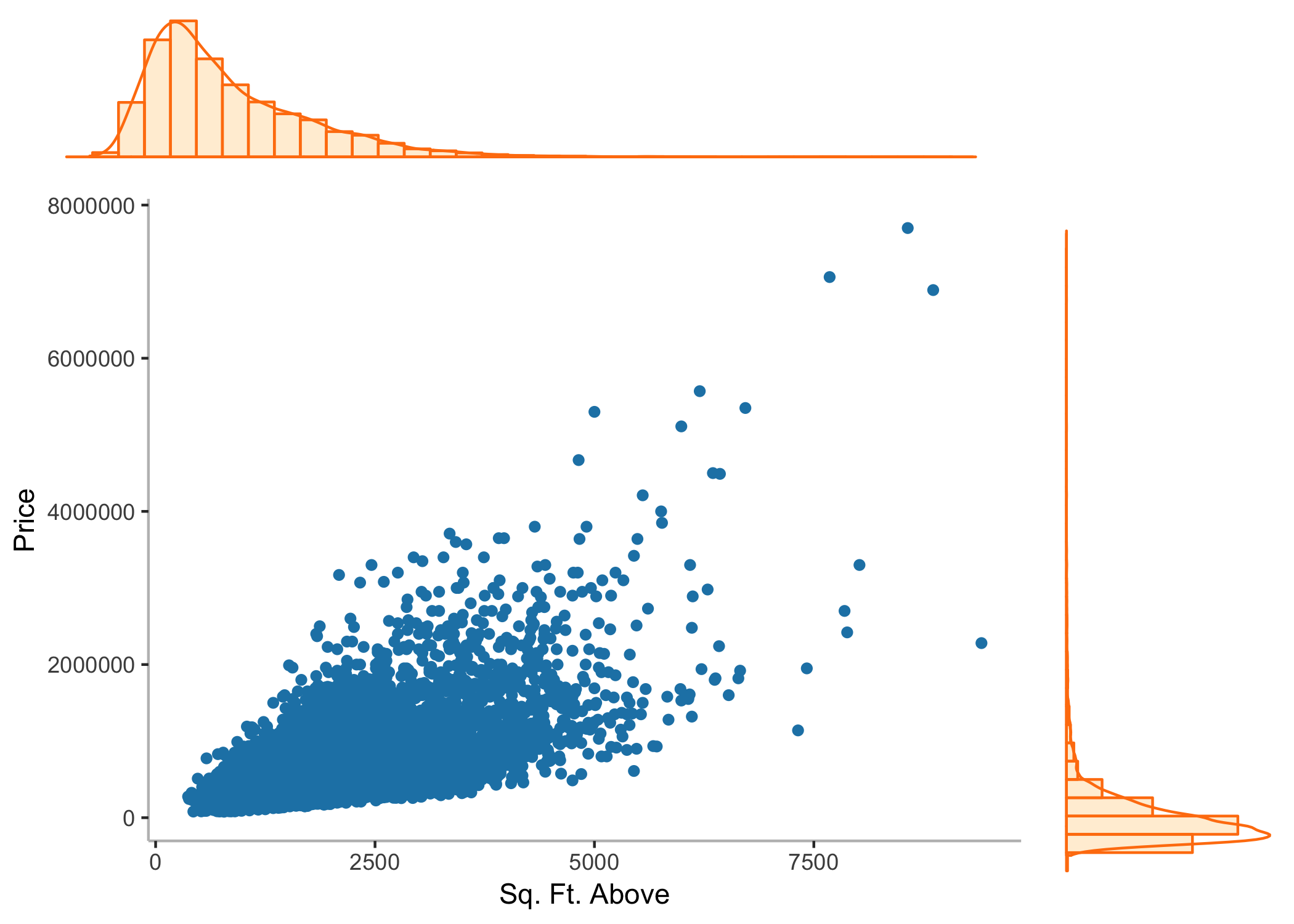
### Scatter/Histogram/Density Plots: Grade vs Price

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$grade, b=kc\_houses$price, lab\_a = "Grade", lab\_b = "Price")



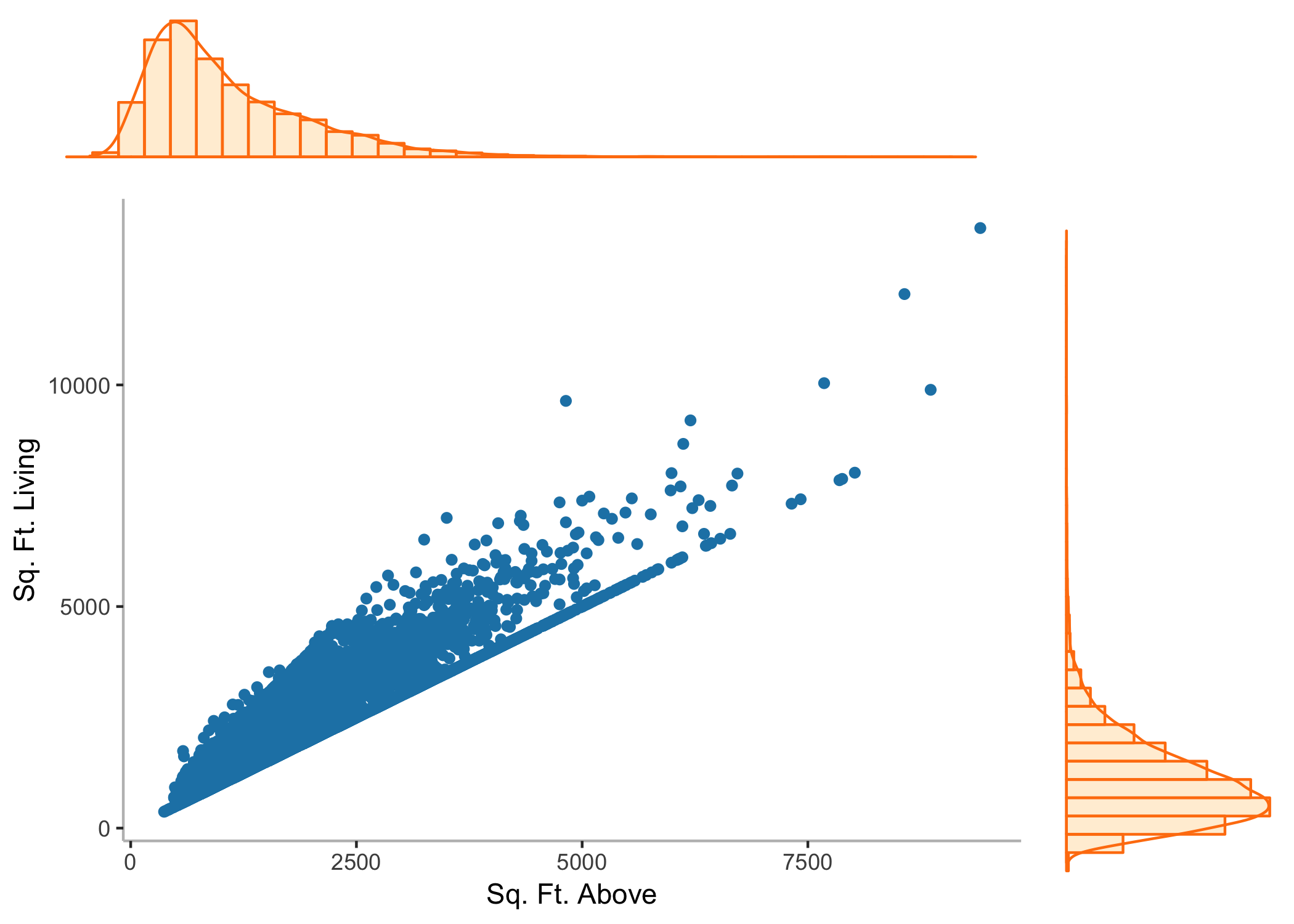
### Scatter/Histogram/Density Plots: Sq. Ft. Above vs Price

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$sqft\_above, b=kc\_houses$price, lab\_a = "Sq. Ft. Above", lab\_b = "Price")



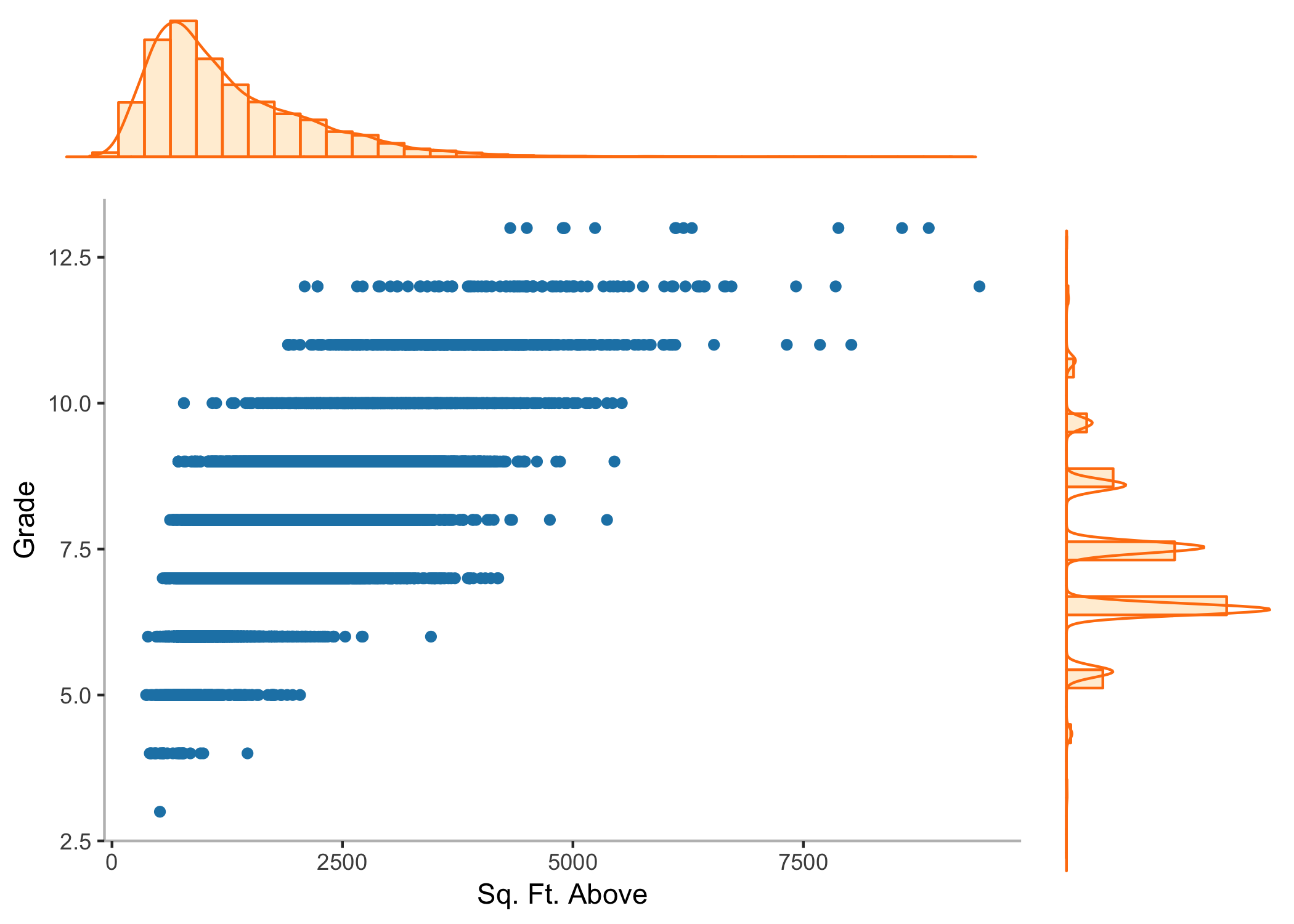
### Scatter/Histogram/Density Plots: Sq. Ft Above vs Sq. Ft. Living

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$sqft\_above, b=kc\_houses$sqft\_living, lab\_a = "Sq. Ft. Above", lab\_b = "Sq. Ft. Living")



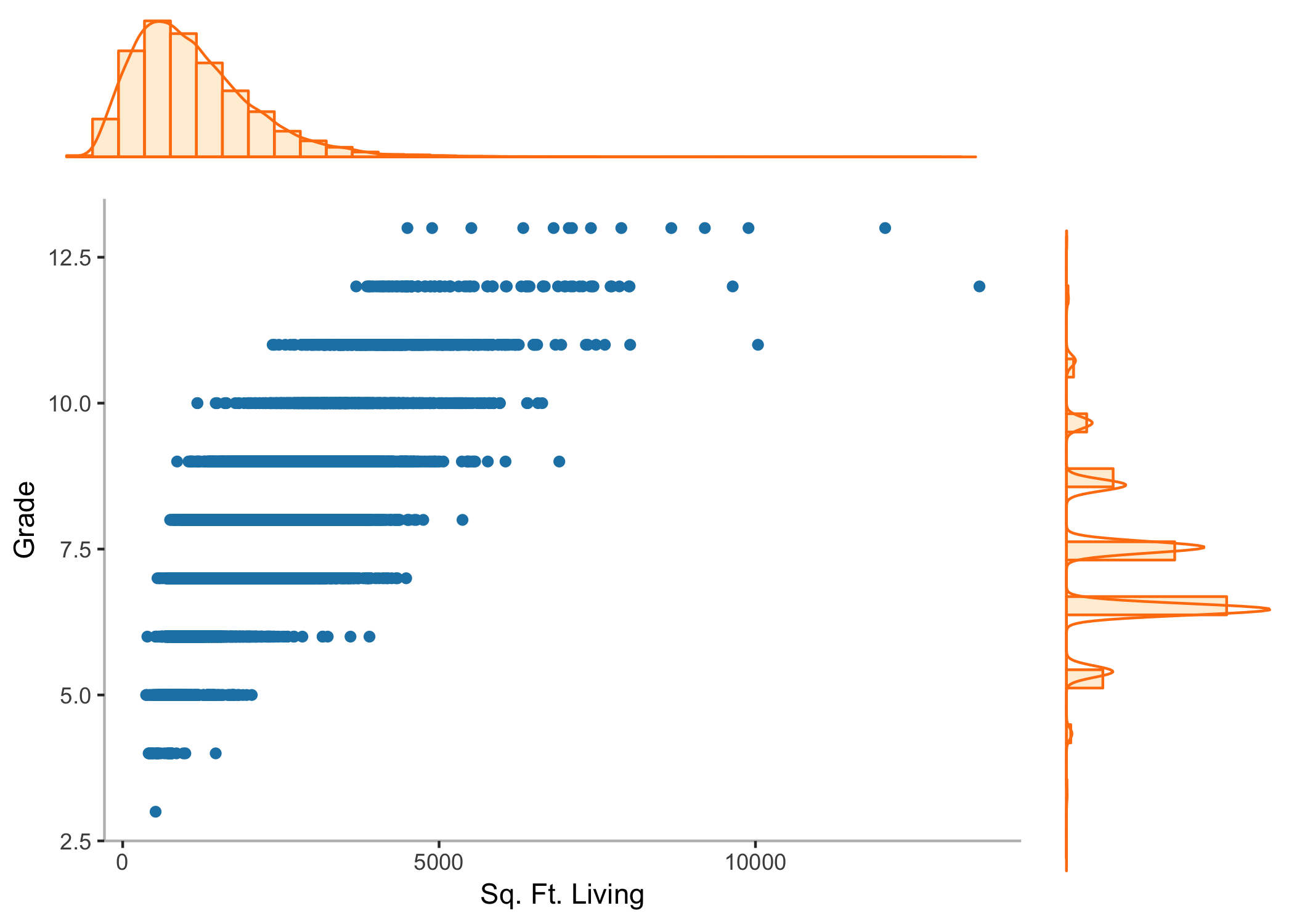
### Scatter/Histogram/Density Plots: Sq. Ft Above vs Grade

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$sqft\_above, b=kc\_houses$grade, lab\_a = "Sq. Ft. Above", lab\_b = "Grade")



### Scatter/Histogram/Density Plots: Sq. Ft Living vs Grade

ggplot\_scatter\_hist\_combo(dat = kc\_houses, a=kc\_houses$sqft\_living, b=kc\_houses$grade, lab\_a = "Sq. Ft. Living", lab\_b = "Grade")



### Fitting a linear regression: Sq. Ft Living vs Price

model<-lm(price~sqft\_living, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ sqft\_living, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1478896 -147583 -24131 106274 4359590   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -43988.892 4410.023 -9.975 <0.0000000000000002 \*\*\*  
## sqft\_living 280.863 1.939 144.819 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 261700 on 21595 degrees of freedom  
## Multiple R-squared: 0.4927, Adjusted R-squared: 0.4927   
## F-statistic: 2.097e+04 on 1 and 21595 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## sqft\_living 1 1435979346411368 1435979346411368 20972  
## Residuals 21595 1478602783996835 68469682056   
## Pr(>F)   
## sqft\_living < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
#vif(model)

### Fitting a linear regression: Grade vs Price

model<-lm(price~grade, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ grade, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -819320 -151846 -36054 98154 6042365   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1061416 12286 -86.4 <0.0000000000000002 \*\*\*  
## grade 209158 1586 131.9 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 273400 on 21595 degrees of freedom  
## Multiple R-squared: 0.4462, Adjusted R-squared: 0.4461   
## F-statistic: 1.74e+04 on 1 and 21595 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## grade 1 1300364813228575 1300364813228575 17396  
## Residuals 21595 1614217317179626 74749586348   
## Pr(>F)   
## grade < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
#vif(model)

### Fitting a linear regression: Sq. Ft. Above vs Price

model<-lm(price~sqft\_above, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ sqft\_above, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -914000 -165893 -41486 109326 5337755   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 59757.111 4737.581 12.61 <0.0000000000000002 \*\*\*  
## sqft\_above 268.668 2.404 111.77 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 292400 on 21595 degrees of freedom  
## Multiple R-squared: 0.3665, Adjusted R-squared: 0.3664   
## F-statistic: 1.249e+04 on 1 and 21595 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## sqft\_above 1 1068107925465501 1068107925465501 12492  
## Residuals 21595 1846474204942702 85504709652   
## Pr(>F)   
## sqft\_above < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
#vif(model)

## Multiple Regression

### Fitting a multiple regression: Sq. Ft. Living + Grade vs Price

model<-lm(price~sqft\_living+grade, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ sqft\_living + grade, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1066199 -138328 -24864 100523 4793864   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -602791.788 13341.637 -45.18 <0.0000000000000002 \*\*\*  
## sqft\_living 184.121 2.872 64.10 <0.0000000000000002 \*\*\*  
## grade 99251.094 2247.784 44.16 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 250600 on 21594 degrees of freedom  
## Multiple R-squared: 0.5347, Adjusted R-squared: 0.5347   
## F-statistic: 1.241e+04 on 2 and 21594 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## sqft\_living 1 1435979346411368 1435979346411368 22865.0  
## grade 1 122444337887608 122444337887608 1949.7  
## Residuals 21594 1356158446109227 62802558401   
## Pr(>F)   
## sqft\_living < 0.00000000000000022 \*\*\*  
## grade < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
print('Variable Inflation Factor')

## [1] "Variable Inflation Factor"

vif(model)

## sqft\_living grade   
## 2.391383 2.391383

### Fitting a multiple regression: Sq. Ft. Living + Sq. Ft. Above vs Price

model<-lm(price~sqft\_living+sqft\_above, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ sqft\_living + sqft\_above, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1505870 -147424 -23666 105379 4338962   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -41266.657 4455.441 -9.262 < 0.0000000000000002 \*\*\*  
## sqft\_living 295.726 4.026 73.448 < 0.0000000000000002 \*\*\*  
## sqft\_above -18.810 4.466 -4.212 0.0000254 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 261600 on 21594 degrees of freedom  
## Multiple R-squared: 0.4931, Adjusted R-squared: 0.4931   
## F-statistic: 1.05e+04 on 2 and 21594 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## sqft\_living 1 1435979346411368 1435979346411368 20988.743  
## sqft\_above 1 1213763626630 1213763626630 17.741  
## Residuals 21594 1477389020370205 68416644455   
## Pr(>F)   
## sqft\_living < 0.00000000000000022 \*\*\*  
## sqft\_above 0.00002542 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
print('Variable Inflation Factor')

## [1] "Variable Inflation Factor"

vif(model)

## sqft\_living sqft\_above   
## 4.313332 4.313332

### Fitting a multiple regression: Grade + Sq. Ft. Above vs Price

model<-lm(price~grade+sqft\_above, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ grade + sqft\_above, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -797336 -149181 -33804 100261 5633593   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -822646.15 14286.16 -57.58 <0.0000000000000002 \*\*\*  
## grade 153695.00 2371.02 64.82 <0.0000000000000002 \*\*\*  
## sqft\_above 103.97 3.36 30.94 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 267500 on 21594 degrees of freedom  
## Multiple R-squared: 0.4697, Adjusted R-squared: 0.4696   
## F-statistic: 9562 on 2 and 21594 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## grade 1 1300364813228575 1300364813228575 18166.6  
## sqft\_above 1 68516724201419 68516724201419 957.2  
## Residuals 21594 1545700592978208 71580095998   
## Pr(>F)   
## grade < 0.00000000000000022 \*\*\*  
## sqft\_above < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
print('Variable Inflation Factor')

## [1] "Variable Inflation Factor"

vif(model)

## grade sqft\_above   
## 2.334516 2.334516

### Fitting a multiple regression: Sq. Ft. Living + Grade + Sq. Ft. Above vs Price

model<-lm(price~sqft\_living+grade+sqft\_above, data = kc\_houses)  
# Coefficent of Determination  
summary(model)

##   
## Call:  
## lm(formula = price ~ sqft\_living + grade + sqft\_above, data = kc\_houses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1120217 -136869 -24076 97184 4758690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -656430.921 13591.272 -48.30 <0.0000000000000002 \*\*\*  
## sqft\_living 234.590 4.039 58.08 <0.0000000000000002 \*\*\*  
## grade 110785.608 2325.608 47.64 <0.0000000000000002 \*\*\*  
## sqft\_above -78.096 4.427 -17.64 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 248800 on 21593 degrees of freedom  
## Multiple R-squared: 0.5413, Adjusted R-squared: 0.5412   
## F-statistic: 8494 on 3 and 21593 DF, p-value: < 0.00000000000000022

# Analysis of Variance  
anova(model)

## Analysis of Variance Table  
##   
## Response: price  
## Df Sum Sq Mean Sq F value  
## sqft\_living 1 1435979346411368 1435979346411368 23193.47  
## grade 1 122444337887608 122444337887608 1977.68  
## sqft\_above 1 19269340667109 19269340667109 311.23  
## Residuals 21593 1336889105442118 61913078564   
## Pr(>F)   
## sqft\_living < 0.00000000000000022 \*\*\*  
## grade < 0.00000000000000022 \*\*\*  
## sqft\_above < 0.00000000000000022 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Variable Inflation Factor  
print('Variable Inflation Factor')

## [1] "Variable Inflation Factor"

vif(model)

## sqft\_living grade sqft\_above   
## 4.797597 2.596616 4.683512