`Міністерство освіти і науки України

Національний університет “Львівська політехніка”

Інститут комп'ютерних наук та інформаційних технологій

Кафедра систем штучного інтелекту



**Лабораторна робота №3**

*З дисципліни:*

**“** **Видуботок великих даних”**

**Виконала:**

Студентка групи КН-318

Яворська М.О.

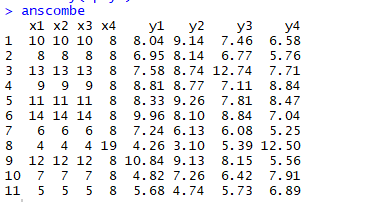
**Перевірила:**

професор Шаховська Н.Б.

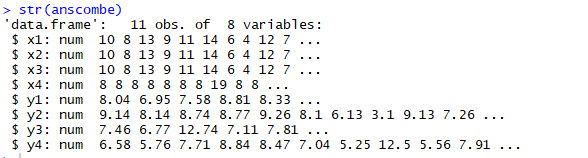
Львів-2019

**Хід роботи**

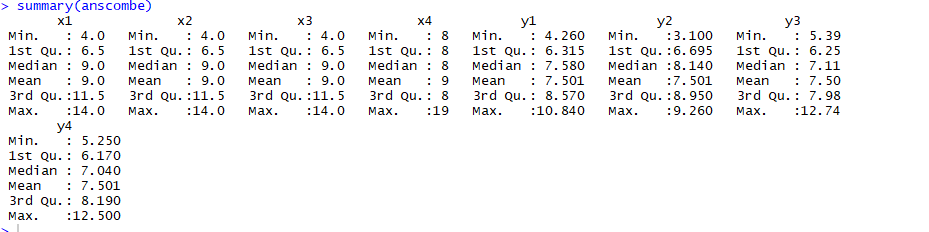
• Explore our data structure: anscombe



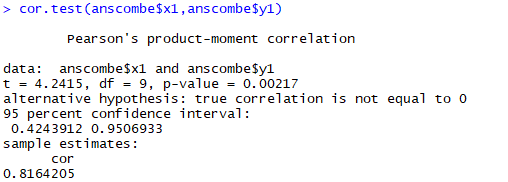
str(anscombe)

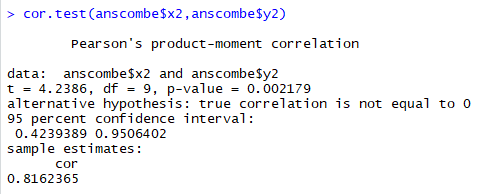


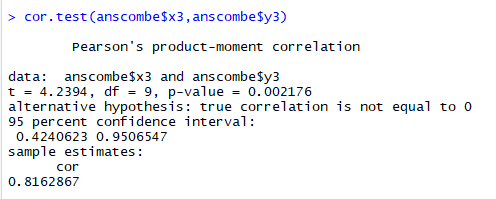
summary(anscombe)

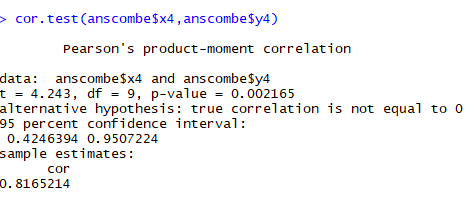


• Calculate the correlation coefficient.

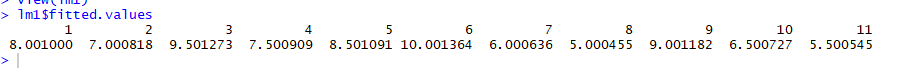




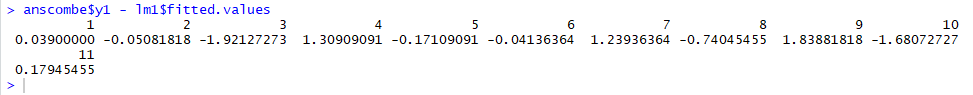




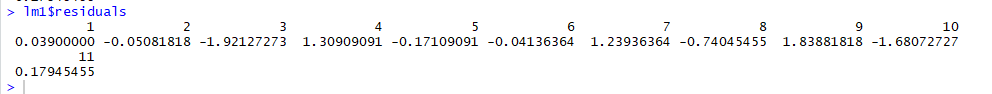
• Find the value ^ (fitted.values) according to our linear regression equation.



• Let’s find the remains. You can do this by subtracting fitted.values real values y1:



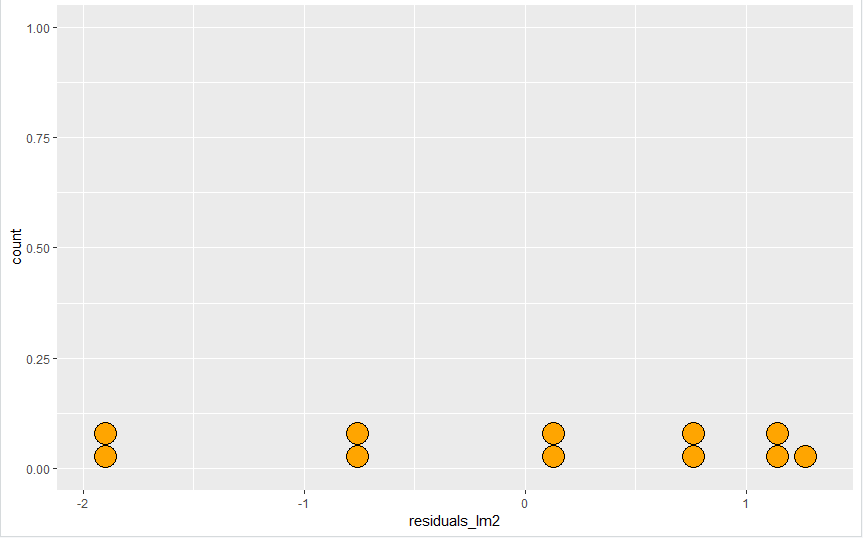
Or use parameter residuals



• The best way for estimating is the distribution of data using bar graphs, but we have just eleven points, so we can use a point schedule.

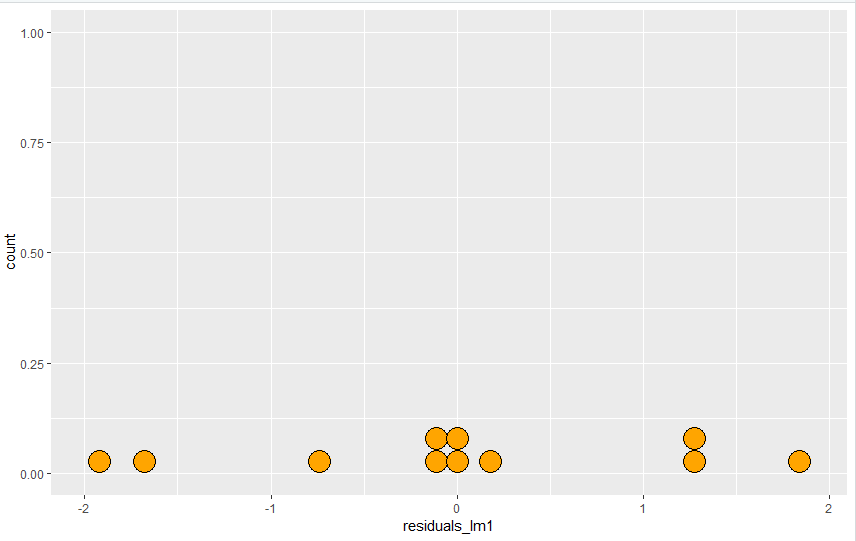
anscombe$residuals\_lm1 <- lm1$residuals

ggplot(anscombe, aes(x = residuals\_lm1)) + geom\_dotplot(fill ="orange")



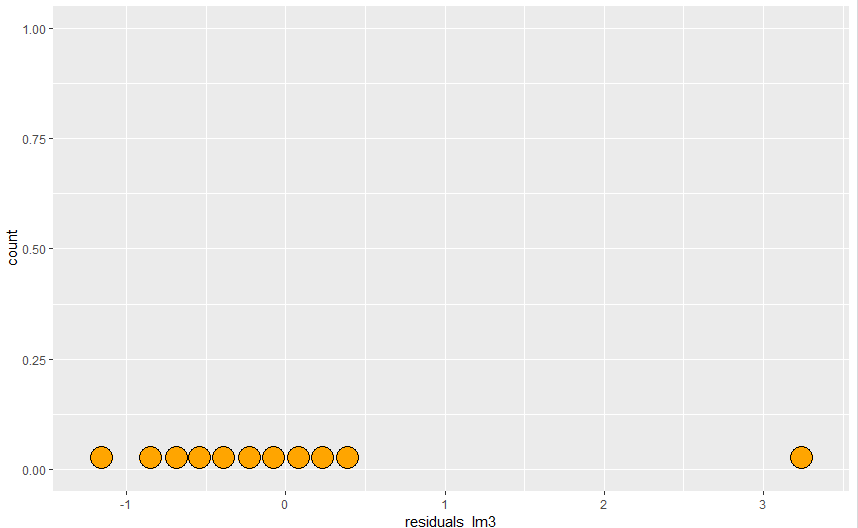
anscombe$residuals\_lm2 <- lm2$residuals

ggplot(anscombe, aes(x = residuals\_lm2)) + geom\_dotplot(fill ="orange")



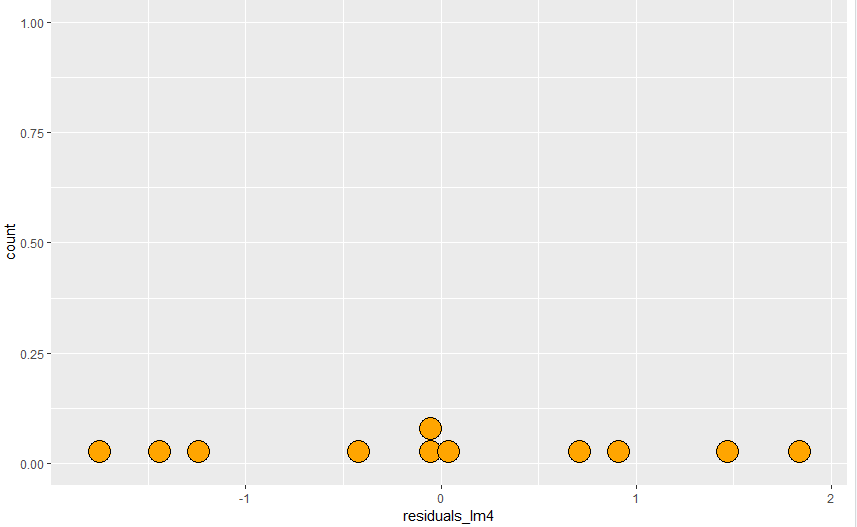
anscombe$residuals\_lm3 <- lm3$residuals

ggplot(anscombe, aes(x = residuals\_lm3)) + geom\_dotplot(fill ="orange")



anscombe$residuals\_lm4 <- lm4$residuals

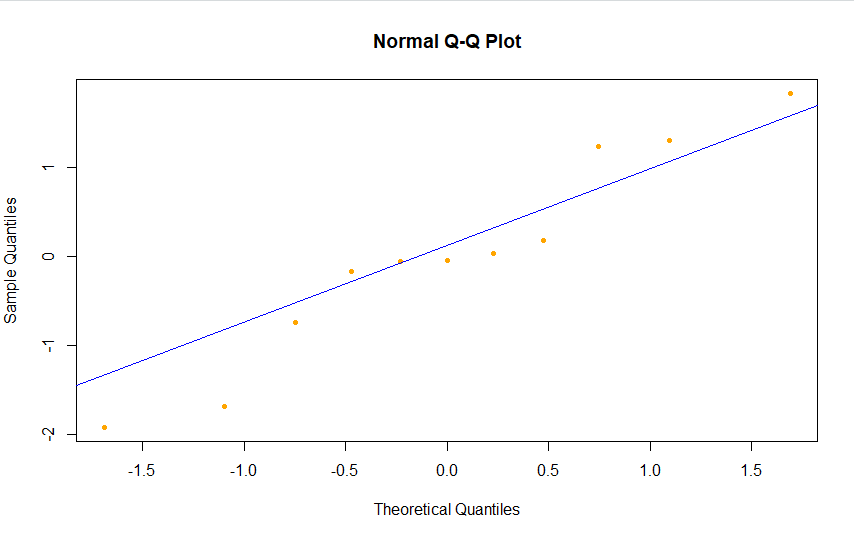
ggplot(anscombe, aes(x = residuals\_lm4)) + geom\_dotplot(fill ="orange")



• To assess the normality of the distribution function will use qqnorm and qqline for (x1, y1):

qqnorm(lm1$residuals, col="orange", pch=20)

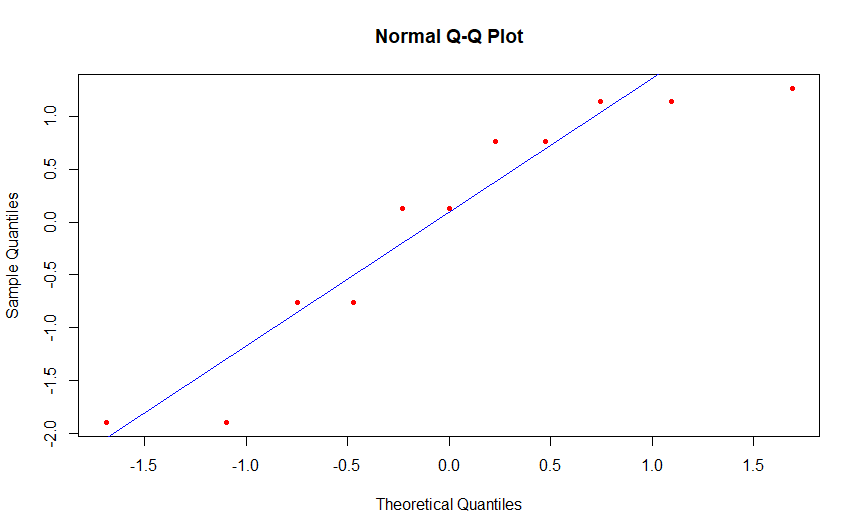
qqline(lm1$residuals, col = "blue")



for (x2, y2):

qqnorm(lm2$residuals, col="red", pch=20)

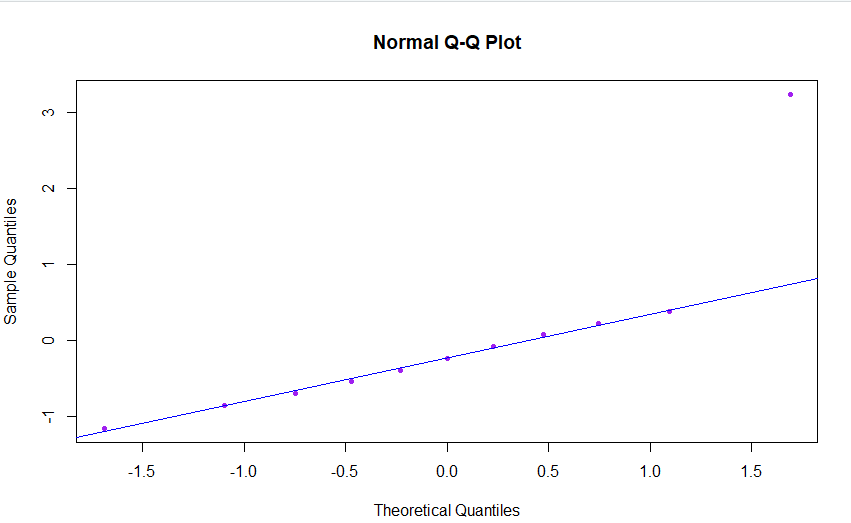
qqline(lm2$residuals, col = "blue")



for (x3, y3):

qqnorm(lm3$residuals, col="purple", pch=20)

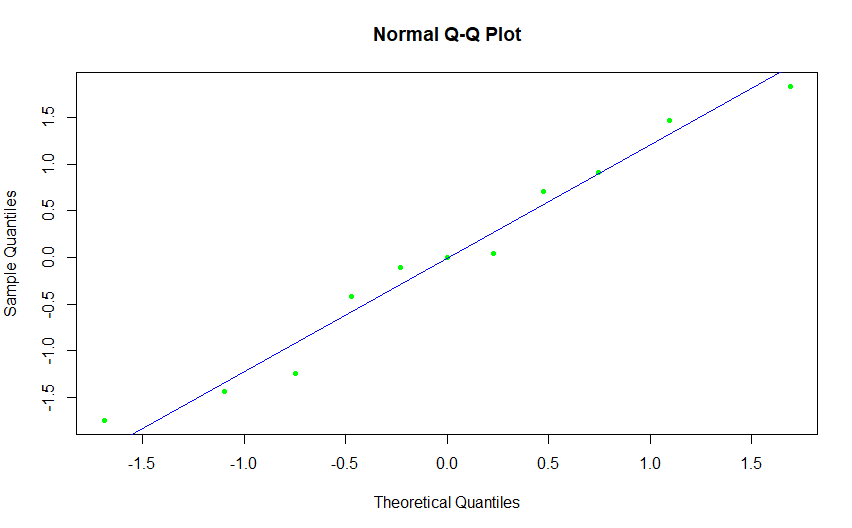
qqline(lm3$residuals, col = "blue")



for (x4, y4):

qqnorm(lm4$residuals, col="green", pch=20)

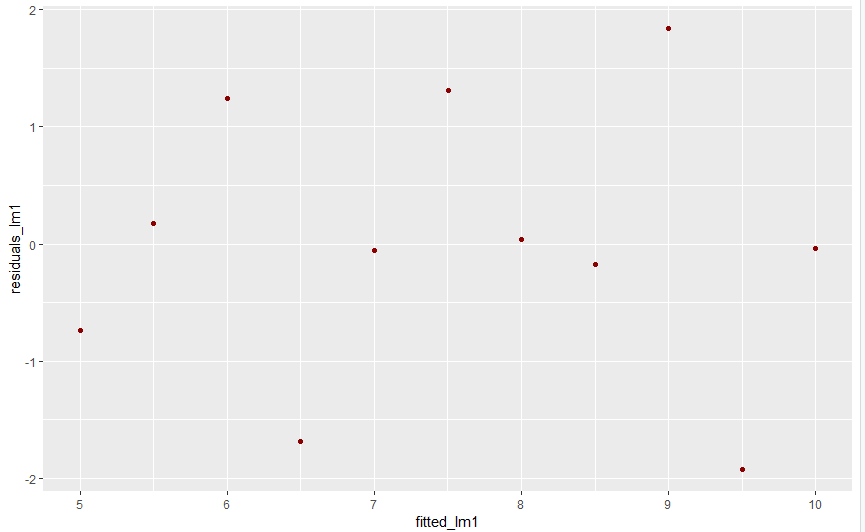
qqline(lm4$residuals, col = "blue")



• Evaluate variability residues: for (x1, y1) condition sustainability of balances is implemented

anscombe$fitted\_lm1 <- lm1$fitted.values

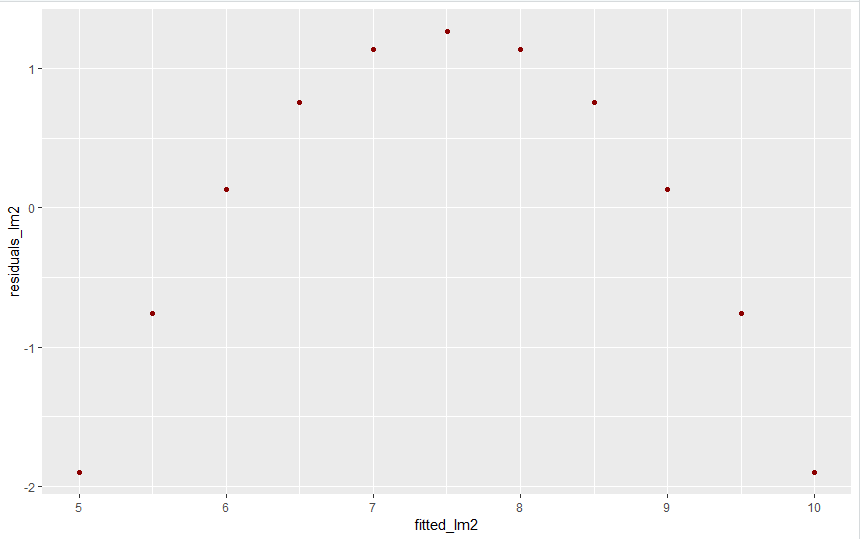
ggplot(data=anscombe, aes(x=fitted\_lm1, y=residuals\_lm1)) + geom\_point(col="darkred")



for (x2, y2) condition sustainability of balances doesn’t implemented.

anscombe$fitted\_lm2 <- lm2$fitted.values

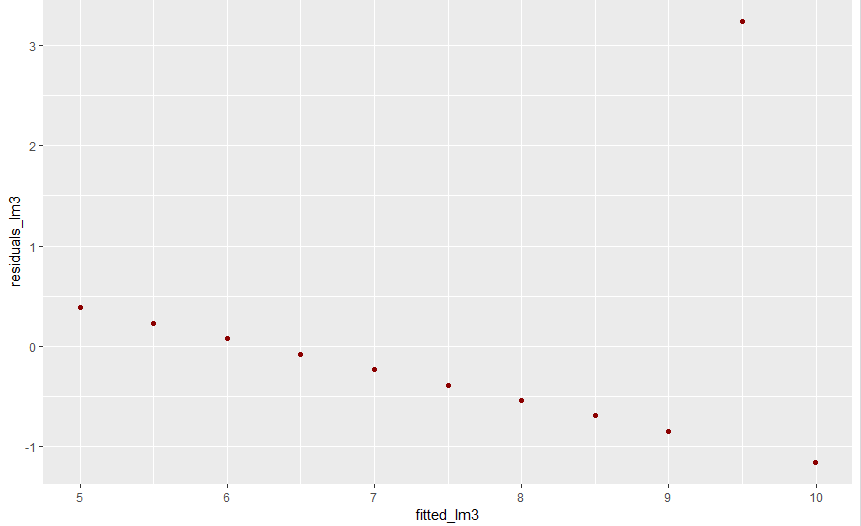
ggplot(data=anscombe, aes(x=fitted\_lm2, y=residuals\_lm2)) + geom\_point(col="darkred")



for (x3, y3) condition sustainability of balances doesn’t implemented.

anscombe$fitted\_lm3 <- lm3$fitted.values

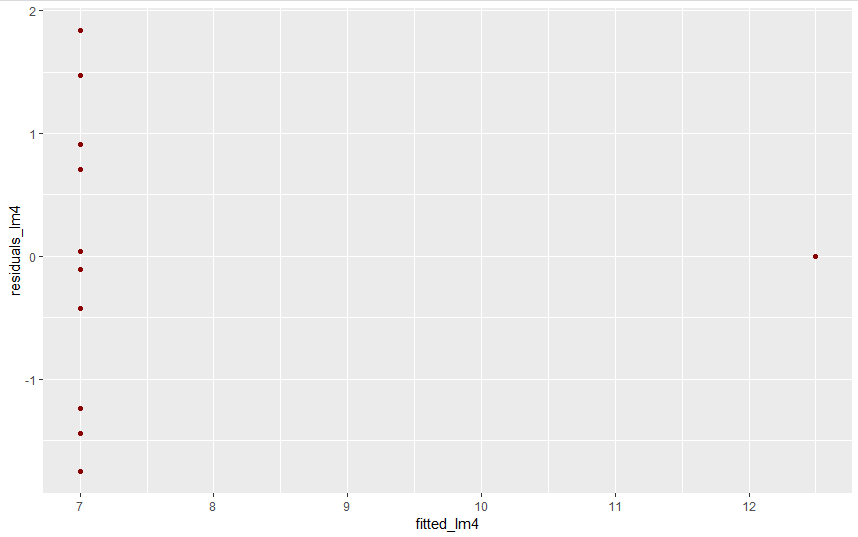
ggplot(data=anscombe, aes(x=fitted\_lm3, y=residuals\_lm3)) + geom\_point(col="darkred")



for (x4, y4) condition sustainability of balances doesn’t implemented.

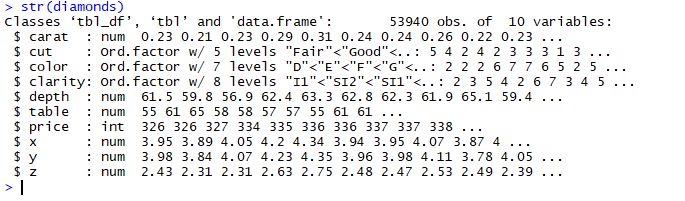
anscombe$fitted\_lm4 <- lm4$fitted.values

ggplot(data=anscombe, aes(x=fitted\_lm4, y=residuals\_lm4)) + geom\_point(col="darkred")



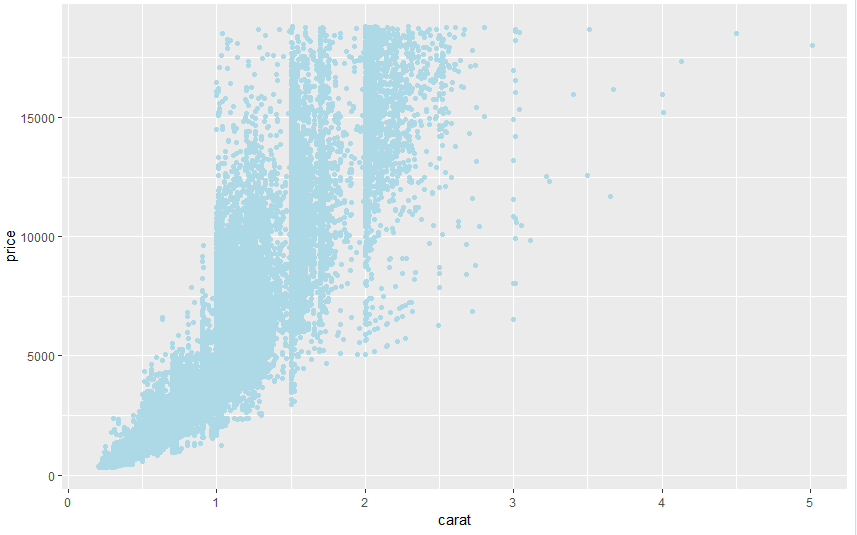
• Let's look at the structure of the data set:

str(diamonds)



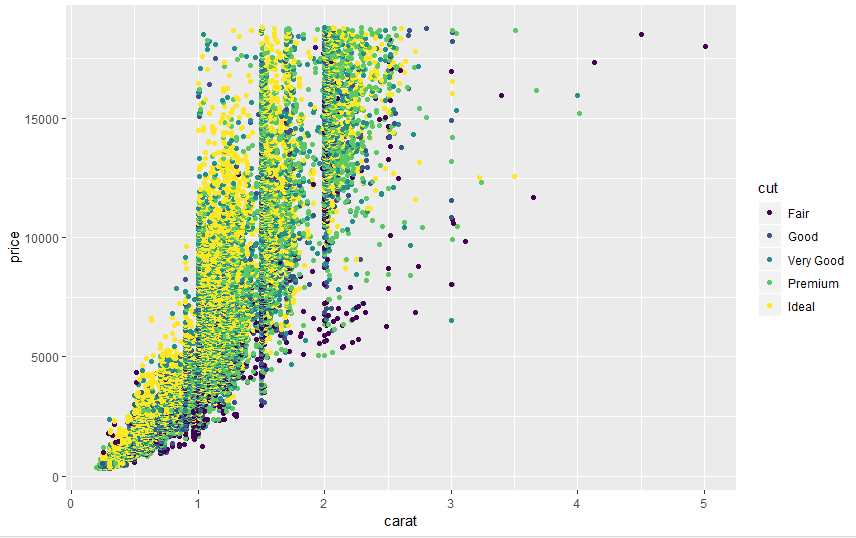
• Find dependence between carat and price

ggplot(data=diamonds, aes(x=carat, y=price)) + geom\_point(col="lightblue")

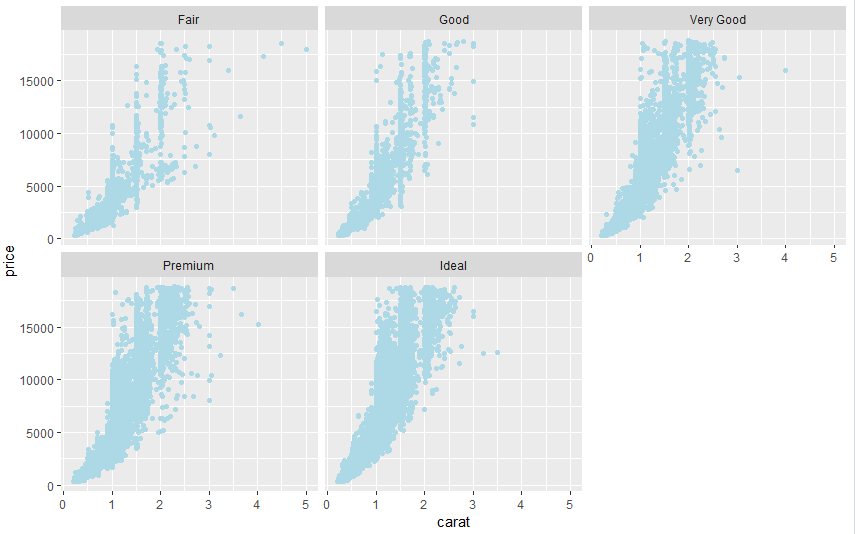


• Find the correlation coeficient carat and price

ggplot(data=diamonds, aes(x=carat, y=price, col=cut)) + geom\_point()

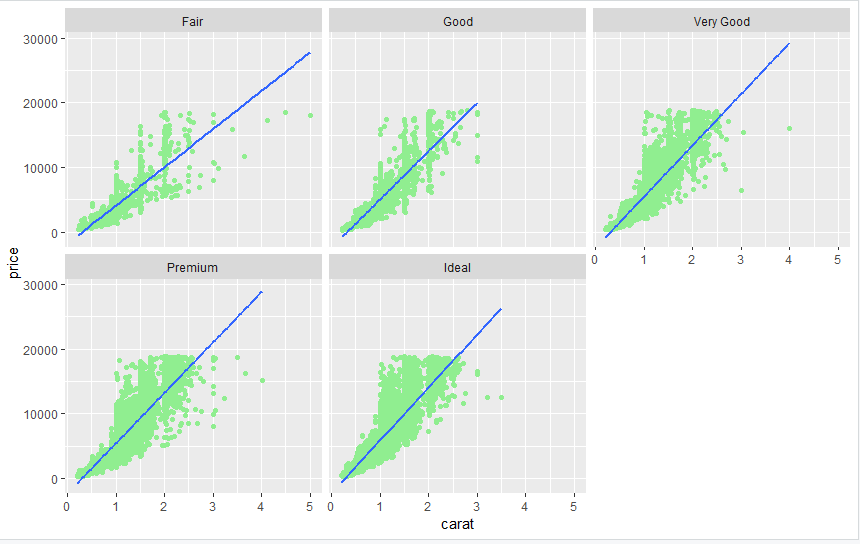


ggplot(data=diamonds, aes(x=carat, y=price)) + geom\_point(col="lightblue") + facet\_wrap(~cut)



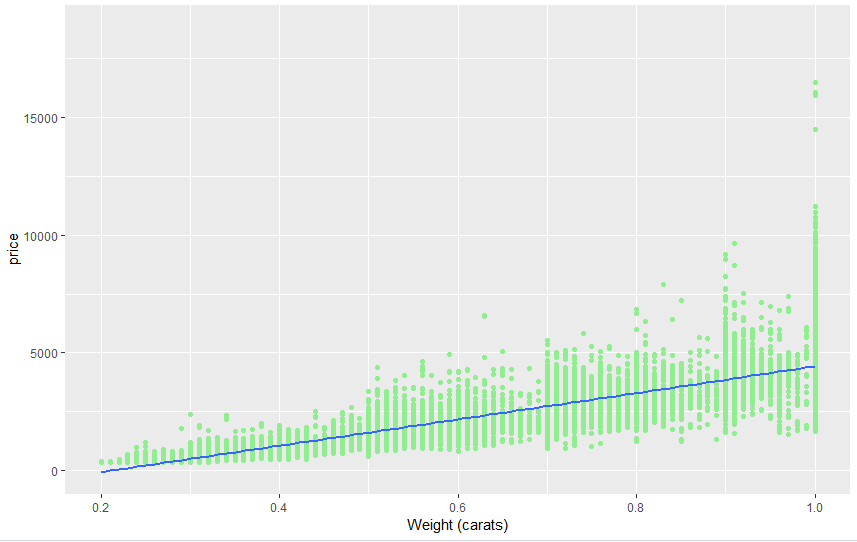
• Add linear regression:

ggplot(data=diamonds, aes(x=carat, y=price)) + geom\_point(col="lightgreen") + geom\_smooth(method="lm", se=FALSE) + facet\_wrap(~cut)



• Find the perfect and acceptable price of processed diamonds weighing 1 carat constructed by linear models.

ggplot(data=diamonds, aes(x=carat, y=price)) + geom\_point(col="lightgreen") + geom\_smooth(method="lm", se=FALSE) + xlab("Weight (carats)") + xlim(0.2, 1)



Around 4500$ is the perfect and acceptable price of processed diamonds weighing 1 carat