Solutions to homework

1. (a) $P(X > 0) = P(X = 1) = 0.3$

   $P(X^2 > 3) = P(X^2 = 4) = 0.5$

   (b) $E(X) = 0 \cdot 0.2 + 1 \cdot 0.3 + (-2) \cdot 0.5 = -0.7$

   $E(X^2) = 0^2 \cdot 0.2 + 1^2 \cdot 0.3 + (-2)^2 \cdot 0.5 = 2.3$

   $E(5X + 10) = 5 \cdot E(X) + 10 = 5 \cdot (-0.7) + 10 = 6.5$

   $Var(X) = E(X^2) - (E(X))^2 = 2.3 - (-0.7)^2 = 1.81$

   $E(X^3) = 0^3 \cdot 0.2 + 1^3 \cdot 0.3 + (-2)^3 \cdot 0.5 = -3.7$

   (c) Mode equals -2

2. (a) Median and mode are equal to 170 as this is Normal distribution.

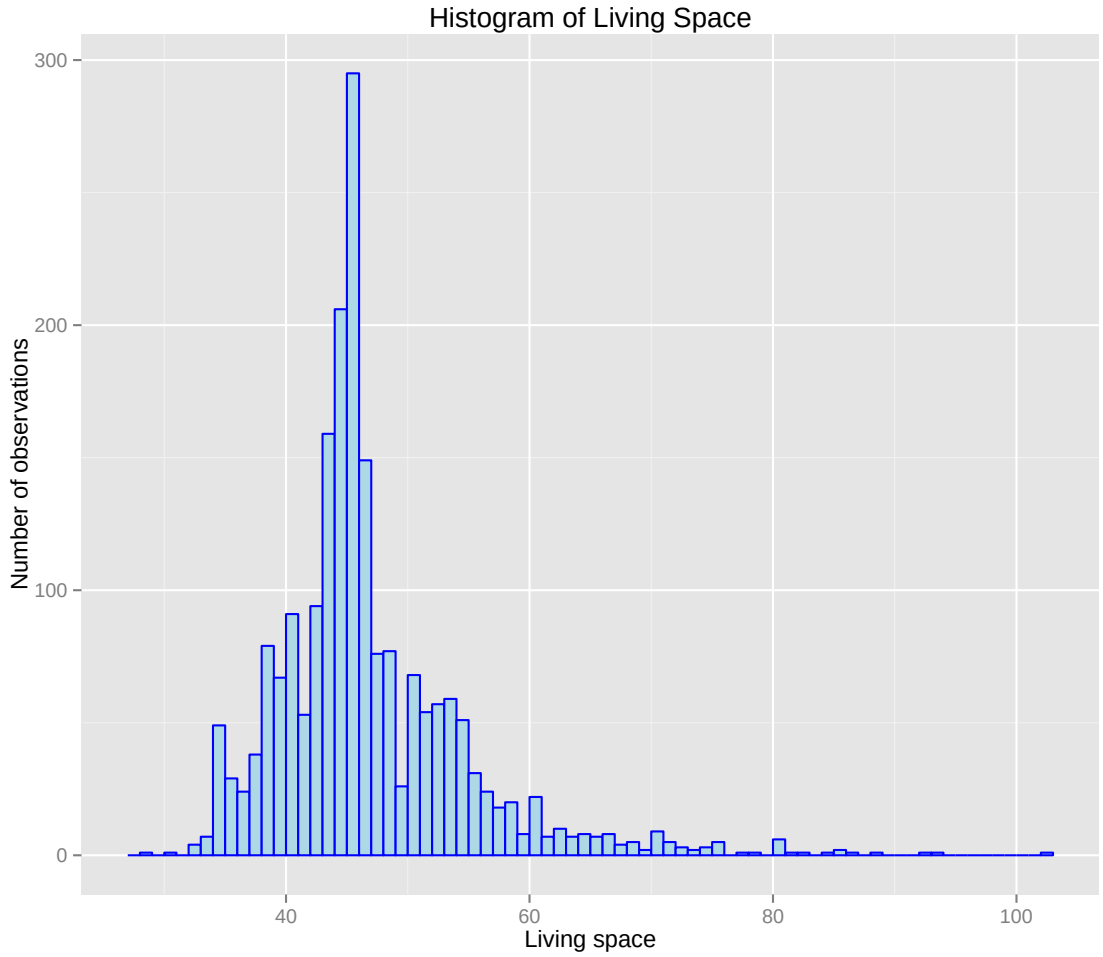
   (b)
```
1 - pnorm(180, mean = 170, sd=5)
```
```
## [1] 0.02275013
```
```
pnorm(165, mean = 170, sd=5)
```
```
## [1] 0.1586553
```
```
pnorm(190, mean = 170, sd=5) - pnorm(165, mean = 170, sd=5)
```
```
## [1] 0.8413131
```

   (c)
```
qnorm(0.95, mean = 170, sd=5)
```
```
## [1] 178.2243
```
```
qnorm(0.10, mean = 170, sd=5)
```
```
## [1] 163.5922
```

3.
```
library(ggplot2)
library(dplyr)
library(texreg)
df <- read.table(file = "flats_moscow.txt",
```

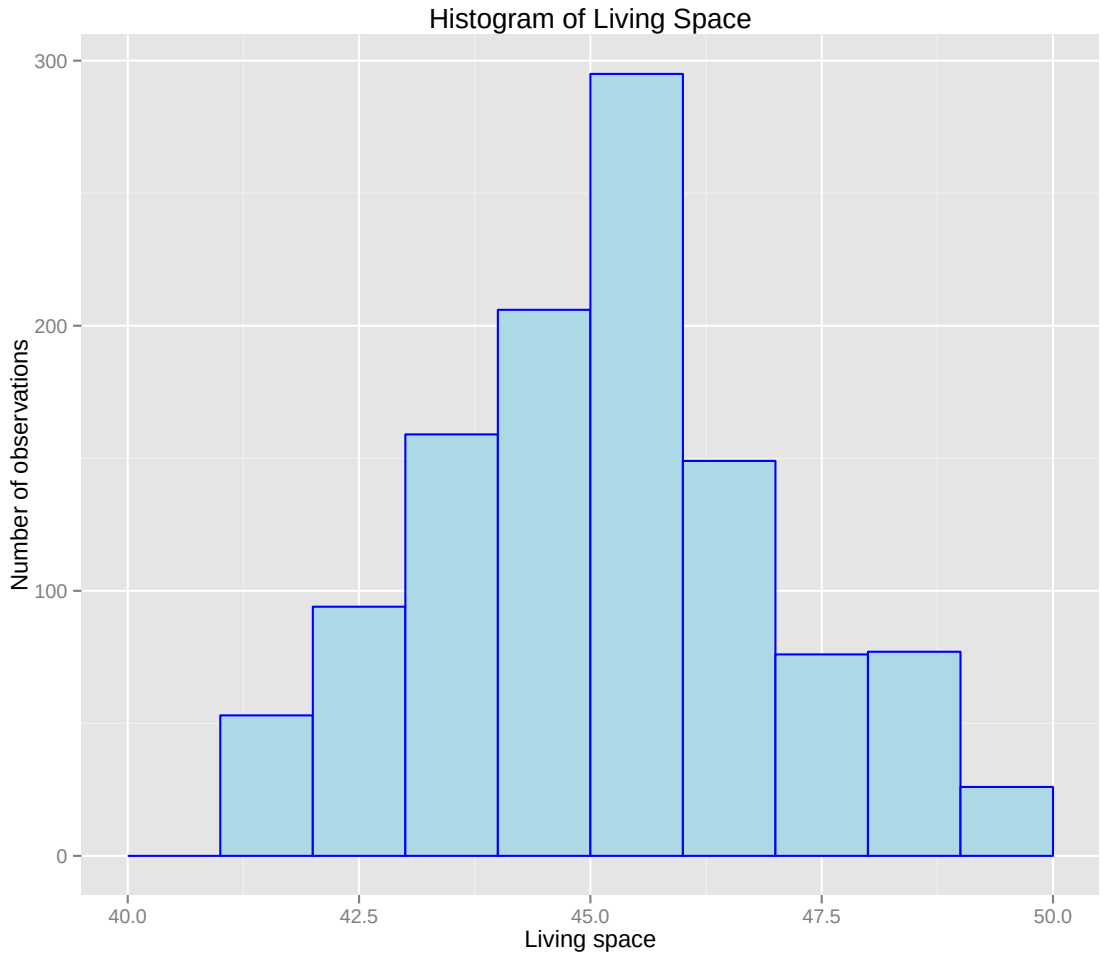
```
                    header = TRUE,
                    dec = ".", sep = "\t")
```

(a)
```
ggplot(df, aes(livesp)) +
  geom_histogram(binwidth=1, fill="lightblue", color="blue") +
  labs(x="Living space",y="Number of observations", title="Histogram of Living
```



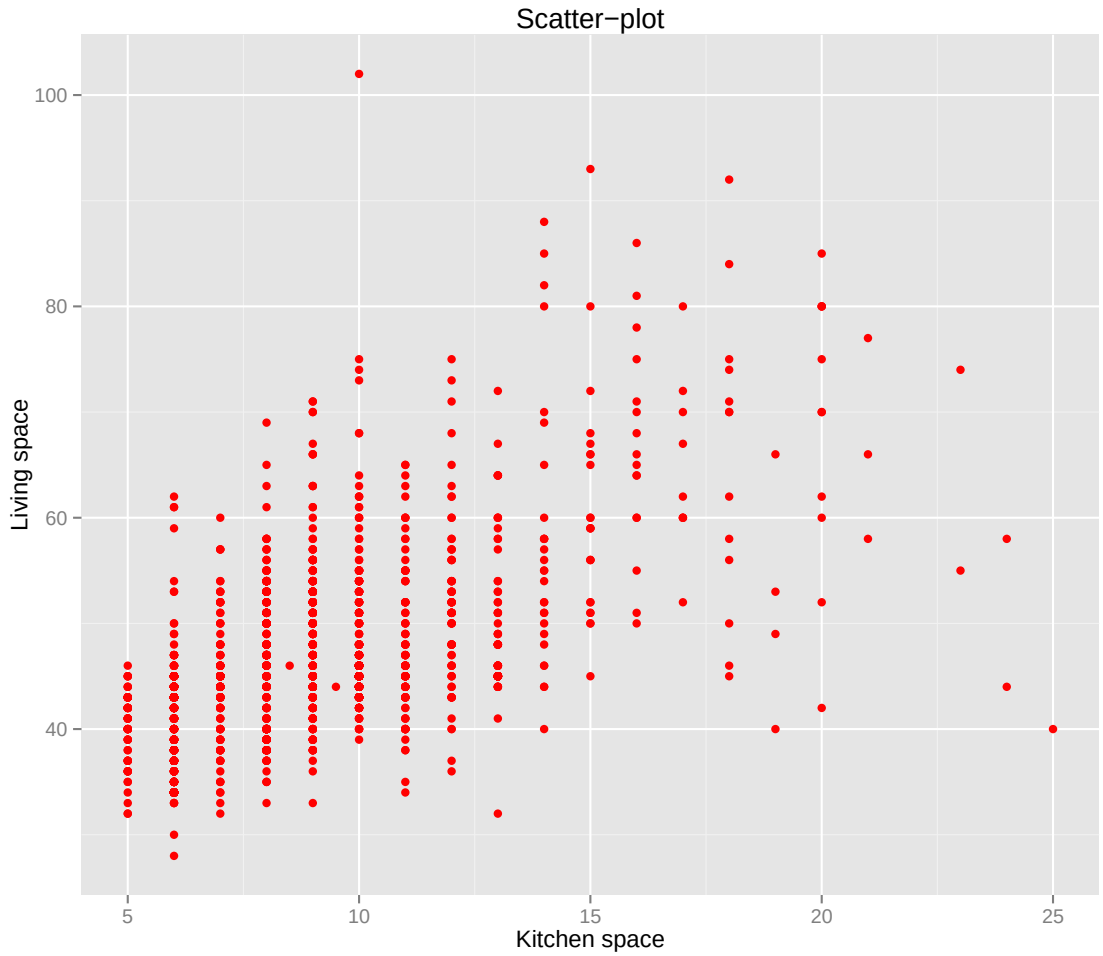
The most frequent living space is 45.

```
df_short <- subset(df, livesp > 40 & livesp < 50)
ggplot(df_short, aes(livesp)) +
  geom_histogram(binwidth=1, fill="lightblue", color="blue") +
  labs(x="Living space",y="Number of observations", title="Histogram of Living
```



(b) On the whole they are positively related.

```
ggplot(df, aes(x=kitsp, y=livesp)) +
  geom_point(color="red") +
  labs(x="Kitchen space",y="Living space", title="Scatter-plot")
```



(c) The size of the biggest kitchen is 25, the smallest living space is 28.

```
df_span <- select(df, c(totsp, livesp, kitsp))
summary(df_span)

##      totsp           livesp           kitsp
##  Min.   : 44.00   Min.   : 28.00   Min.   : 5.000
##  1st Qu.: 62.00   1st Qu.: 42.00   1st Qu.: 7.000
##  Median : 73.50   Median : 45.00   Median : 9.000
##  Mean   : 73.08   Mean   : 46.34   Mean   : 8.899
##  3rd Qu.: 79.00   3rd Qu.: 50.00   3rd Qu.:10.000
##  Max.   :192.00   Max.   :102.00   Max.   :25.000
```

(d) The correlation is positive and is slightly greater than 0.5, which means that the linear connection is quite strong and when kitchen space increases, living space increases as well. This is logical as you won't have a big kitchen in a tiny flat.

```r
cor(df$kitsp, df$livesp)

## [1] 0.5735282
```

4. Our model is

$$price_i = \beta_0 + \beta_{livesp} \cdot livesp_i + \beta_{kitsp} \cdot kitsp_i + \beta_{brick} \cdot brick_i + \varepsilon_i$$

   (a) Testing each coefficient for significance separately.

   $H_0 : \beta_i = 0$

   $H_A : \beta_i \neq 0$

   Criterion:

   - P-value > significance level, do not reject H_0 (the coefficient is not significant, does not influence the price)
   - P-value < significance level, reject H_0 (the coefficient is significant, influences the price)

```r
model <- lm(data = df, price ~ livesp + kitsp + brick)
# summary(model)
```

```r
#texreg(model,single.row = TRUE)
```

	Model 1
(Intercept)	−86.43 (4.40)***
livesp	3.44 (0.12)***
kitsp	5.57 (0.33)***
brick	14.78 (1.65)***
R^2	0.60
Adj. R^2	0.60
Num. obs.	2040
RMSE	33.00

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

For each coefficient P-value is less than 2e-16, which is smaller than any reasonable significance level (1%, 5%, 10%), so we reject the null hypothesis that each beta is equal to zero. So each of coefficients is significant and influences the price.

(b) Testing the whole regression for significance.

$H_0 : \beta_{livesp} = 0, \beta_{kitsp} = 0, \beta_{brick} = 0, \beta_0 = 0$

H_A : At least one beta is not equal to 0

Criterion:

- P-value > significance level, do not reject H_0 (which means that the model is bad)
- P-value < significance level, reject H_0 (which means that the model is contains at least one variable influencing the price)

P-value for the whole model is 2.2e-16, which is smaller than any reasonable significance level (1%, 5%, 10%), so we reject the null hypothesis that all betas are equal to zero. It means that at least one coefficient is significant, it influences the price. So model is not bad.

(c) Brick houses are priced higher than the ones that are not brick ones as the coefficient is positive (14.78) and significant.

5.
```r
model1 <- lm(data = df, price ~ livesp + kitsp + brick)
model2 <- lm(data = df, price ~ brick)
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ brick
## Model 2: price ~ livesp + kitsp + brick
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2038 5114310
## 2   2036 2217757  2   2896553 1329.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Comparing two models.

We can test 1%, 5% and 10% significance levels (standard).

$H_0 : \beta_{livesp} = 0, \beta_{kitsp} = 0$

H_a : At least one beta is not equal to 0

Criterion:

- P-value > significance level, do not reject H_0 (which means that restricted model works well, we do not need to include living space and kitchen space in the regression)
- P-value < significance level, reject H_0 (which means that living space and kitchen space should be included in the regression)

(b) Model 1 is unrestricted (long). Model 2 is restricted (short).

(c) P-value is less than 2.2e-16.

(d) P-value is less than any reasonable significance level (1%, 5%, 10%) so we reject the null hypothesis that β_{livesp} and β_{kitsp} are equal to zero as P-value is smaller than any reasonable significance level. The restricted model is worse than the unrestricted one, we should include living space and kitchen space in our model.