In the name of God

Home Work #1
Statistical Inference

Amin Asadi
810196410

Problem 1.

a	
type of the study	experimental
explanatory variables	drinking ayahuasca or placebo
response variable	level of depression
establish causal relationships?	Yes, because random assignment was used

b	
type of the study	observational
explanatory variables	religious affiliations
response variable	Life length
establish causal relationships?	No, because random assignment was not used.(Actually It is not possible, since you can not force a person to have or have not religious affiliations)

c	
type of the study	experimental

explanatory variables	Using viral medication or not, i.e. using placebo)
response variable	Usefulness on bacterial infection
establish causal relationships?	Yes, because random assignment was used

d	
type of the study	observational
explanatory variables	Being flight attendant or not
response variable	incidence rate of cancer
establish causal relationships?	No, because random assignment was not used.

e	
type of the study	observational
explanatory variables	used device(desktop or mobile)
response variable	Amount of time spent on the website
establish causal relationships?	No, because random assignment was not used.

f	
type of the study	experimental
explanatory variables	Receiving daily notification or not
response variable	Amount of logging of workouts

establish causal relationships?	Yes, because random assignment was used

---

Problem 2.

a. This is simple random sampling. 5 employees are selected randomly (since the ball associated to each of them is selected randomly).
b. This is stratified sampling. Each model of trucks is a strata. Then some trucks are selected randomly from each strata.
c. This is systematic sampling. First an initial sample(50th passenger) is selected from near the beginning and then every 50th sample is selected thereafter. Thus the method is systematic sampling.
d. This is cluster sampling. Each box is a cluster and all the t-shirts are selected from the chosen clusters(boxes).
e. This is simple random sampling. Each student identification number is an observation and because the computer is generating the identification numbers randomly, the method is simple random sampling.
f. This is stratified sampling. Each of the university levels is a strata. 25 observations are selected randomly from each strata(university levels).
g. This is systematic sampling. First an initial sample(10th cake) is selected from near the beginning and then every tenth sample(cake) is selected thereafter. Thus the method is systematic sampling.
h. This is stratified sampling. Each of the teams is a strata. 3 observations(athletes) are selected randomly from each strata(sport team).

Problem 3.

a.

Population is all the seniors at Riverview High School. Sample is the 10 volunteer seniors of that school.

- One problem is voluntary response bias : this sample might not be representative of the population because the volunteers are more likely to be dissatisfied with the lunch offerings at the school(voluntary bias may exist).
- Another problem might be the sample size. If the population size(number of all the seniors at that school) is very big in comparison to 10, then sample size is small and might not be representative of the whole population.

b.

Population is all the names in the city phone directory. Sample is the randomly selected 75 names.

- Problem is non-response bias. Since non-random 40% of whom she called didn't answer, and these 40% non-respondents could actually change the results, there is non-response bias here.
- Anecdotal Evidence: If the population size(number of all the seniors at that school) is very big in comparison to 10, then sample size is small.

c.

Population is all the vehicles that pass over the bridge. Sample is the first 10th vehicle.

- Anecdotal Evidence: The problem is that the size of the sample set is very small(=1). Very small sample size might not be representative of this large population.

d.

Population is all the food which Ali serves in his restaurant. Sample is the 70 selected entrees.

- Ali’s experiment results will be reliable only if the temperature inside the restaurant is kept the same all day. Otherwise he must repeat his experiment for different temperatures inside the restaurant, because it affects the temperature of foods.
- If the time between when Ali serves the foods and the time the food is given to customers is long enough, he must repeat the experiment for different types of foods because the watery foods take more time to be cold than other foods(Specific heat capacity of water is high).
- Entrees must be representative of all the foods he serves.

Problem 4:

a.

The pool excluded people too young to vote is not a potential source of bias. Because apparently the aim of this survey is to find information about the people who are likely to vote in the upcoming election and find out if they are going to vote or not and since too young people are not eligible to vote for the upcoming election, they are not influential on the results and excluding them will not cause bias.

I. Non-response bias
II. Convenience-Sample bias
III. Non-response bias
IV. -
V. Response-bias because some people have lied and this misinformation causes bias and if the number of such responses are high, then the results of the survey are not reliable.

---

b.

There is response bias in this survey. The reason is that because having the phone out during the class is not allowed and is considered forbidden, an student is more likely to deny ever having the phone out during the class(i.e. is probably afraid of the consequences of it!). Hence the 10% positive response is not probably reliable. This is the reason why the principal found out that more than 25%(more than 10%) of the students had had the phones out during the class.

c.

- Some people don't even check their emails so don't answer that email.
- These types of emails are probable to be marked as spam by the email-service-provider of some customers and hence some customers might not even see that email.
- Some people may be suspicious about this email and for security considerations don't answer this type of emails.

Problem 5.

a.

Yes, there is confusion. variable. The confounding variable is a student's studying hours and quality of being hardworking. Here the explanatory variable is drinking coffee and the response variable is exam scores. A hard worker student may sleep less than others and sleep late at night or wake up soon in the morning and hence drink more coffee. A hard worker student is more likely to get better scores on the exams. So this is not the coffee that results in higher scores, but the quality of being hard working.

b.

- One potential confounding variable is the cost of the sunglasses which may have increased from 2017 to 2018 and hence people have bought them less.

- Another company may have produced similar sunglasses with lower cost so most people have bought that company's sunglasses more and hence the sales of this company has decreased.

c.

The confounding variable is exercise. Here the explanatory variable is being a coal miner or farmworker(i.e. job) and the response variable is lung capacity. Because farm-workers are likely to exercise more and exercise affects lung capacity, so exercise is correlated with both explanatory and response variables and thus is a confounding variable.

d.

- They should placebo so that the members of each group don't know whether they are using the old device or the new device because knowing this may affect their blood pressure and act as a confounding variable.
- Hydration must be performed for both groups not just for those using the new device because it may affect blood pressures and create bias in results.
- If some of the people have heart problems, we have to use stratified sampling to make sure that the number of those people are the same in both groups.

Problem 6.

value	occurrence
0.00	7
0.04	17
0.08	5

0.12	5
0.16	2
0.20	3
0.24	1

a.

- mean = $\frac{0\times7 + 0.04\times17 + 0.08\times5 + 0.12\times5 + 0.16\times2 + 0.20\times3 + 0.24\times1}{40} \simeq 0.071$

- Because there are 40 samples, median is the mean of 20th and 21th sample → median = $\frac{0.04 + 0.04}{2} = 0.04$
- Mode is the most frequent sample which is 0.04
- Range is the difference between min and max samples which is 0.24 - 0.00 = 0.24

- interquartile range is the difference between first and third quartiles(25th and 75th percentiles).

  First quartile = $\frac{0.04 + 0.04}{2} = 0.04$

  Third quartile = $\frac{0.11 + 0.11}{2} = 0.11$

  IQR = Third quartile - First quartile = 0.11 - 0.04 = 0.07

b.

Because the mentioned article says that the proportion of the people who exercise daily is 6%, the null hypothesis assumes that it is the same for the teenagers(6%). So the null hypothesis is:

Null Hypothesis(H0): Proportion of the teenages who exercise daily is equal to 6%.

Because the researcher thinks that the mentioned proportion is higher for the teenagers, it means that his alternative hypothesis suggests a proportion higher than 6%. So the alternative hypothesis is:

Alternative Hypothesis(HA): Proportion of the teenages who exercise daily is larger than 6%.

c.

p-value is the probability of the mentioned proportion in the sample being more than 20% →

$P(proportion\ of\ people\ who\ exercise\ daily \geq 20\% \mid H_0\ is\ True)$

As we can see in the dot plot, there are 4 dots for 0.20 and 1 dot for 0.24. All the other dots are for values less than 0.20. So overall there are 4 samples out of 40 samples for which proportion of the people who exercise daily is >= 6%.

→ p-value ≃ $\frac{4}{40} = 10\%$

d.

P-value says that the probability of the mentioned proportion in the sample being more than 20% is equal to 10%. We should compare the p-value with a threshold. If p-value is more than this threshold we will reject H_0 in favor of alternative hypotheses. Otherwise we can not reject H_0.

Problem 7.

a.

I. If we assume that “Pearson Correlation" is meant here, the answer is “True” because this type of correlation means linear association between variables. Otherwise it is False.

II. True, correlation has a range between -1 and 1. When two variables have a correlation equal to -1 then they are strongly negatively correlated. When correlation is equal to 0 then they are not correlated at all. When the correlation is 1 then they are strongly positively correlated. 0.95 is very close to 1 and hence there is a strong correlation between the two variables.

III. False, since correlation doesn't imply causation. So we can not say smoking causes the increase of lung cancer even if they are strongly correlated. There may be a confounding variable that causes high correlation or even a strong correlation may occur by chance.

b.

I. True. Since the number of texts sent each day is positively correlated with the average credit card debt(0.45) so by definition of correlation, as the number of texts sent each day increases, average credit card debt increases.

II. False. Since correlation doesn't imply causation and hence we can not infer this statement.

III. True. Correlation has a range between -1 and 1 and the closer the absolute value of a correlation is close to 1, the stronger it is. So the correlation between average credit card debt and the number of texts sent each day is stronger. (|0.45| is greater than |-0.3|)

c.

I. There is a linear negative association

II. Again a linear negative association

III. No, because if they were independent, correlation would be zero but as it is apparent from the scatter plot, the correlation between X and Y is not non-zero.

Problem 8.

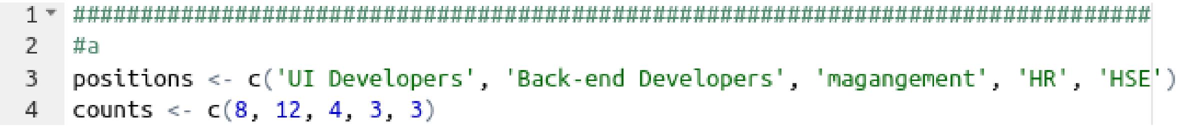
a. Median is the best because the distribution is skewed(left-skewed) and median is not highly sensitive to outliers and rarely-occurring points. We should not use “mean” because it is sensitive to outliers and rarely-occurring points. Also mode is not good enough because it doesn’t show the center sufficiently.

b. IQR is the best because the distribution is skewed(left-skewed). If we use range, since it is super sensitive to extreme values it will not be a good spread measure. Standard deviation would be sensitive to outliers and rarely-occurring points and hence it will not be a good choice, too. But IQR is a good measure here because it is not highly sensitive to outliers and rarely-occurring points and ideal for skewed distributions.

Problem 9:

a.

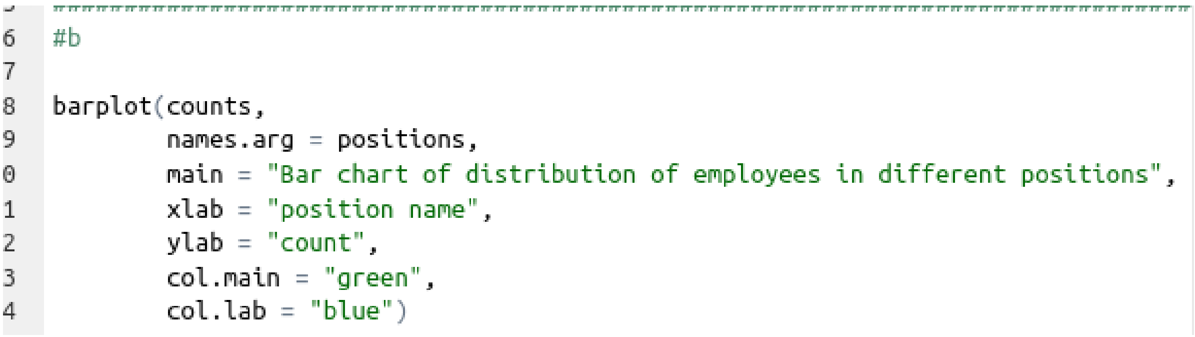
Code:



```
################################################################################
#a
positions <- c('UI Developers', 'Back-end Developers', 'magangement', 'HR', 'HSE')
counts <- c(8, 12, 4, 3, 3)
```

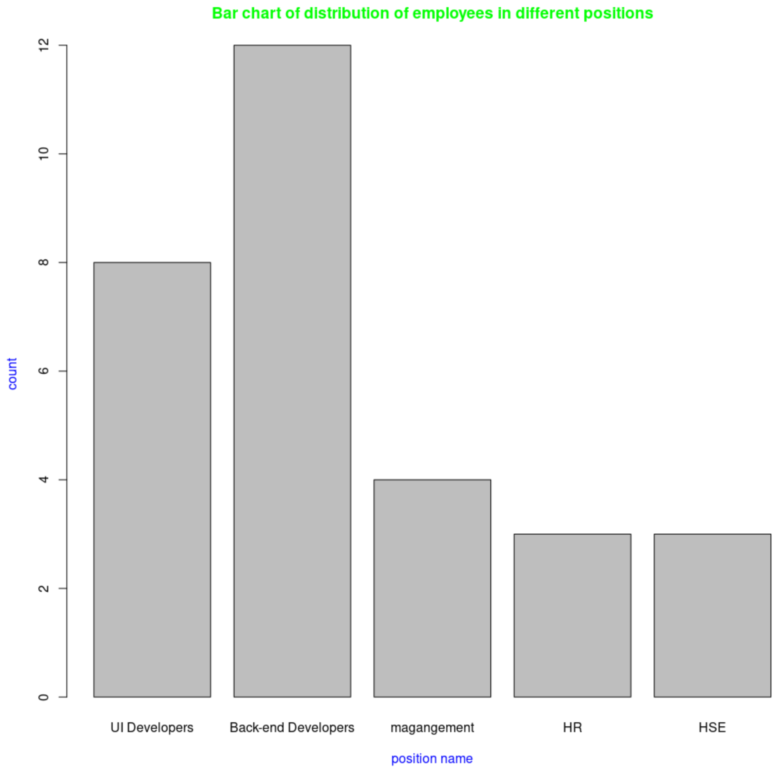
b.

code:



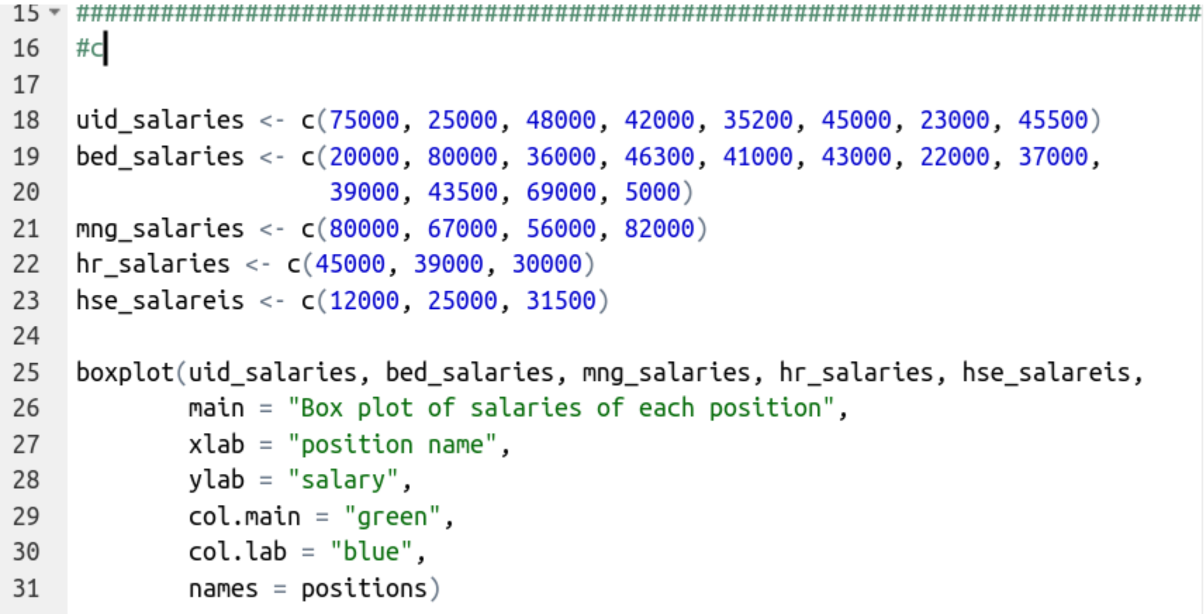
```
##################################################################################
#b

barplot(counts,
        names.arg = positions,
        main = "Bar chart of distribution of employees in different positions",
        xlab = "position name",
        ylab = "count",
        col.main = "green",
        col.lab = "blue")
```

b: result:



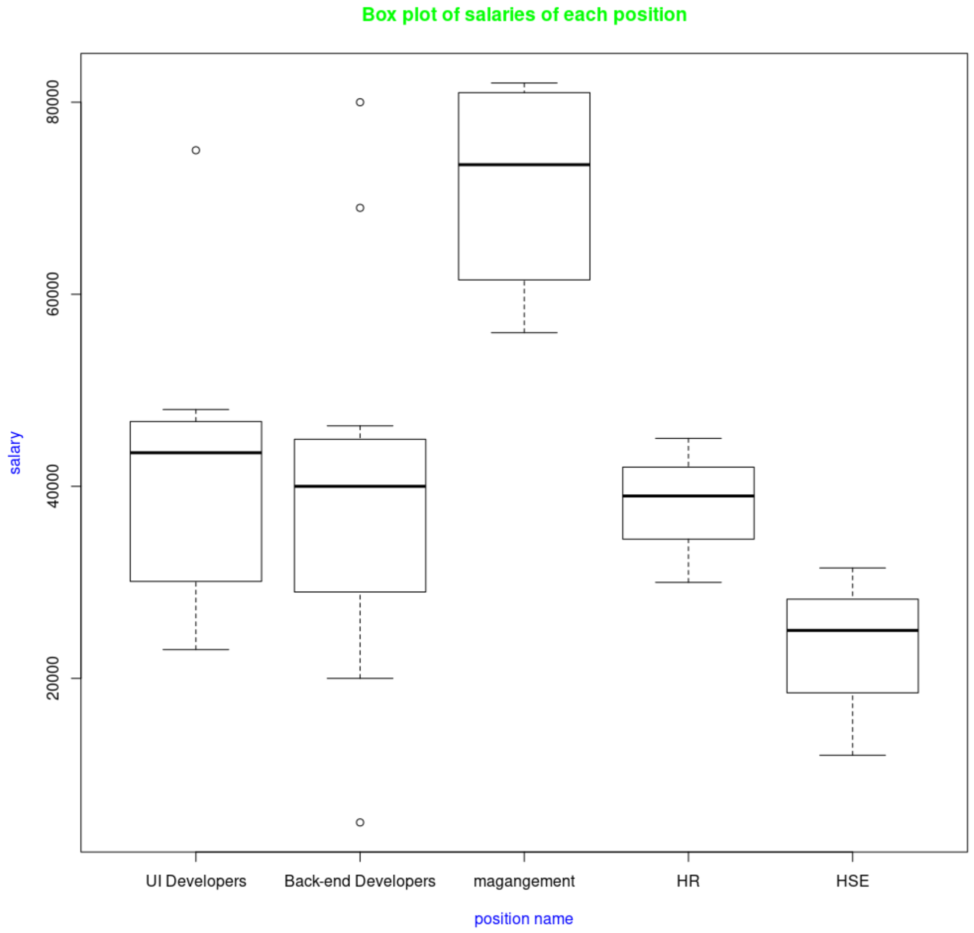
c.

code:



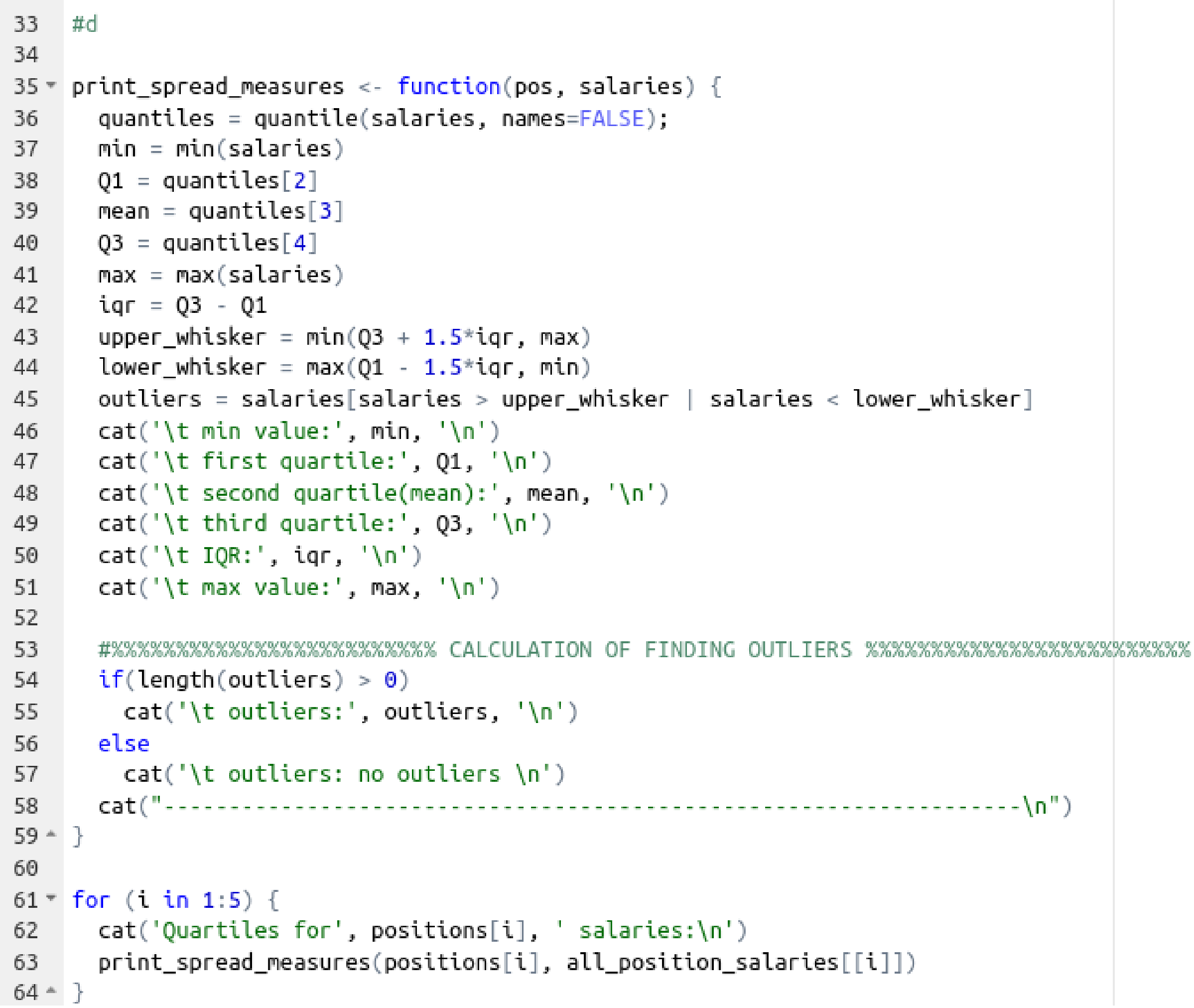
```
################################################################################
#c

uid_salaries <- c(75000, 25000, 48000, 42000, 35200, 45000, 23000, 45500)
bed_salaries <- c(20000, 80000, 36000, 46300, 41000, 43000, 22000, 37000,
                  39000, 43500, 69000, 5000)
mng_salaries <- c(80000, 67000, 56000, 82000)
hr_salaries <- c(45000, 39000, 30000)
hse_salareis <- c(12000, 25000, 31500)

boxplot(uid_salaries, bed_salaries, mng_salaries, hr_salaries, hse_salareis,
        main = "Box plot of salaries of each position",
        xlab = "position name",
        ylab = "salary",
        col.main = "green",
        col.lab = "blue",
        names = positions)
```

c: Result:


Box plot of salaries of each position
salary
80000
60000
40000
20000
UI Developers
Back-end Developers
magangement
HR
HSE
position name

d.

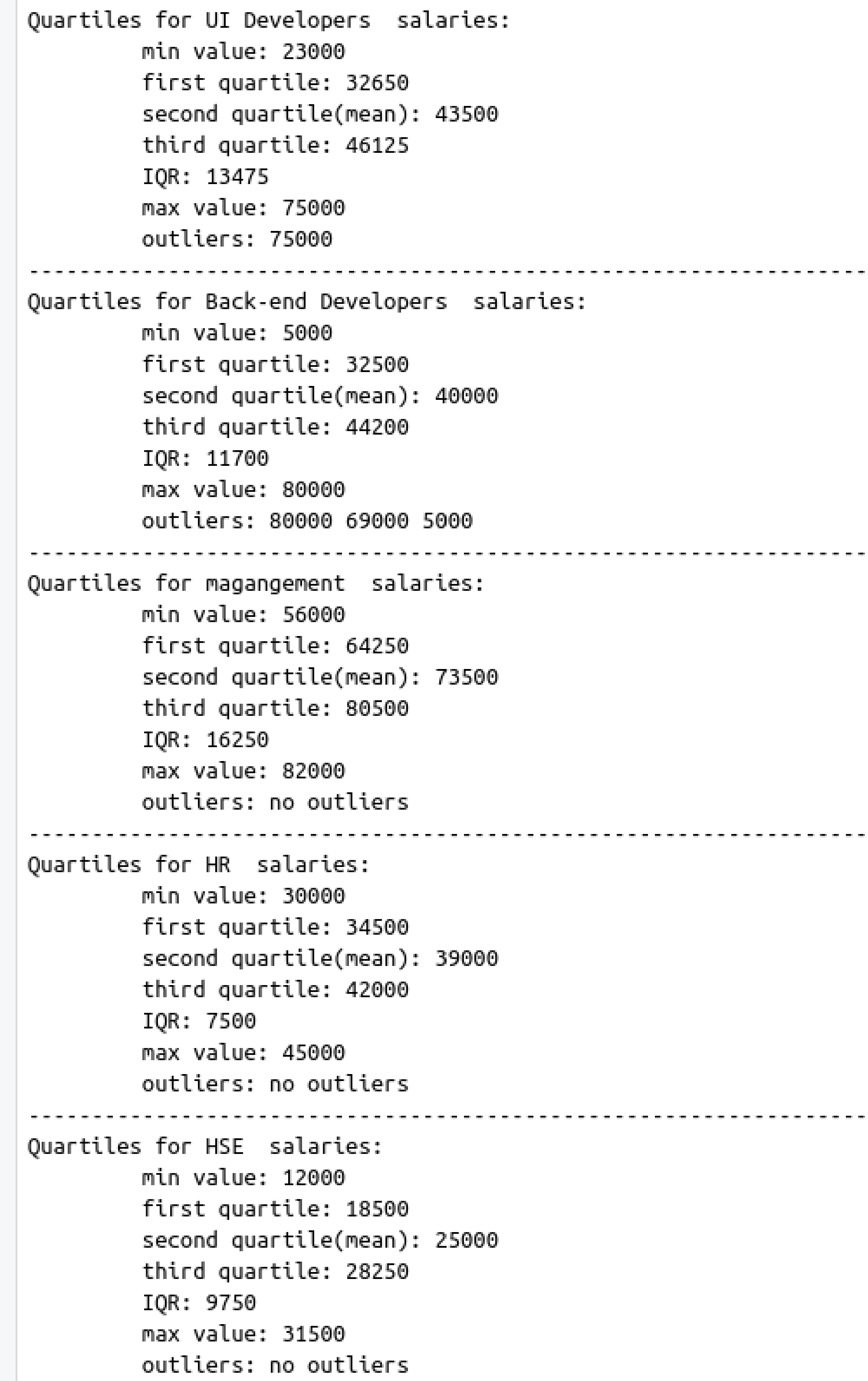
code:



```r
#d

print_spread_measures <- function(pos, salaries) {
  quantiles = quantile(salaries, names=FALSE);
  min = min(salaries)
  Q1 = quantiles[2]
  mean = quantiles[3]
  Q3 = quantiles[4]
  max = max(salaries)
  iqr = Q3 - Q1
  upper_whisker = min(Q3 + 1.5*iqr, max)
  lower_whisker = max(Q1 - 1.5*iqr, min)
  outliers = salaries[salaries > upper_whisker | salaries < lower_whisker]
  cat('\t min value:', min, '\n')
  cat('\t first quartile:', Q1, '\n')
  cat('\t second quartile(mean):', mean, '\n')
  cat('\t third quartile:', Q3, '\n')
  cat('\t IQR:', iqr, '\n')
  cat('\t max value:', max, '\n')

  #%%%%%%%%%%%%%%%%%%%%%%%%% CALCULATION OF FINDING OUTLIERS %%%%%%%%%%%%%%%%%%%%%%%%%
  if(length(outliers) > 0)
    cat('\t outliers:', outliers, '\n')
  else
    cat('\t outliers: no outliers \n')
  cat("------------------------------------------------------------------\n")
}

for (i in 1:5) {
  cat('Quartiles for', positions[i], ' salaries:\n')
  print_spread_measures(positions[i], all_position_salaries[[i]])
}
```

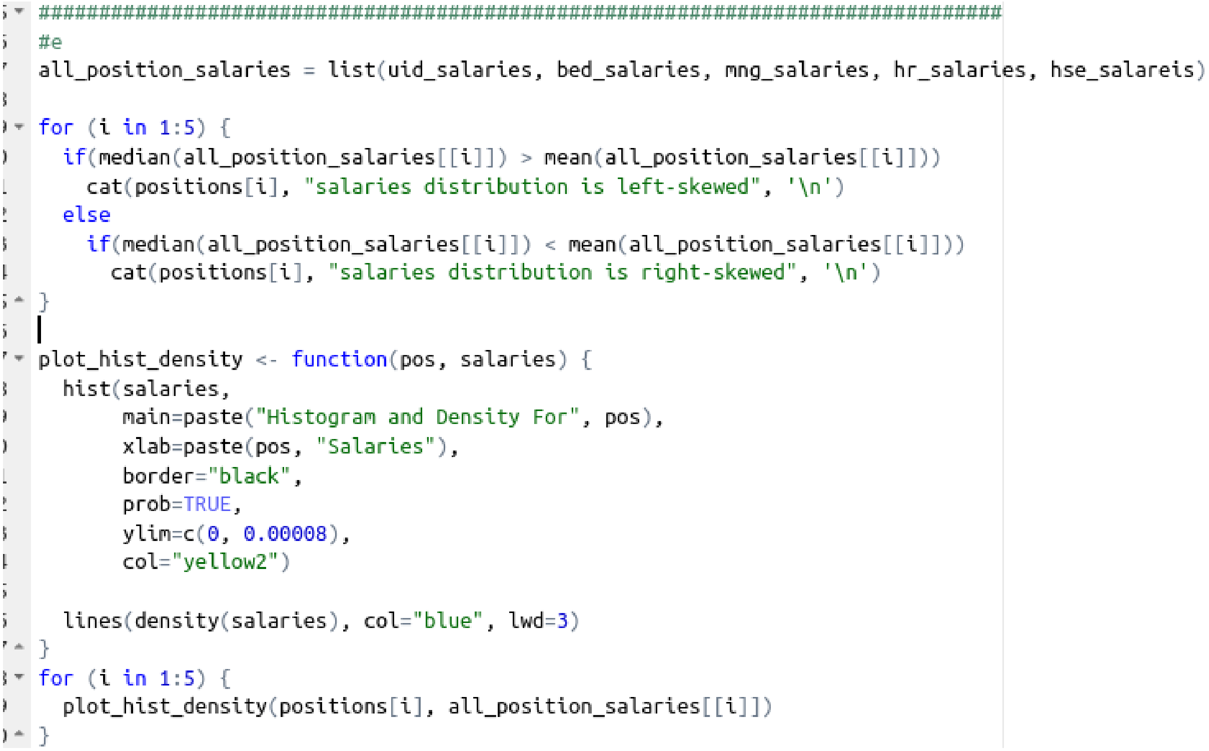
Result:

As we can see only UI Developers and Backend Developers salaries have outliers.



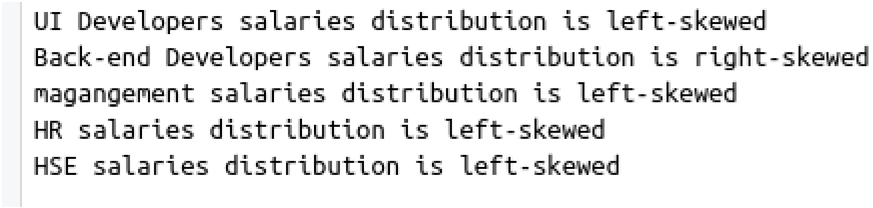
```
Quartiles for UI Developers  salaries:
         min value: 23000
         first quartile: 32650
         second quartile(mean): 43500
         third quartile: 46125
         IQR: 13475
         max value: 75000
         outliers: 75000
------------------------------------------------------------------
Quartiles for Back-end Developers  salaries:
         min value: 5000
         first quartile: 32500
         second quartile(mean): 40000
         third quartile: 44200
         IQR: 11700
         max value: 80000
         outliers: 80000 69000 5000
------------------------------------------------------------------
Quartiles for magangement  salaries:
         min value: 56000
         first quartile: 64250
         second quartile(mean): 73500
         third quartile: 80500
         IQR: 16250
         max value: 82000
         outliers: no outliers
------------------------------------------------------------------
Quartiles for HR  salaries:
         min value: 30000
         first quartile: 34500
         second quartile(mean): 39000
         third quartile: 42000
         IQR: 7500
         max value: 45000
         outliers: no outliers
------------------------------------------------------------------
Quartiles for HSE  salaries:
         min value: 12000
         first quartile: 18500
         second quartile(mean): 25000
         third quartile: 28250
         IQR: 9750
         max value: 31500
         outliers: no outliers
```

e.
code:



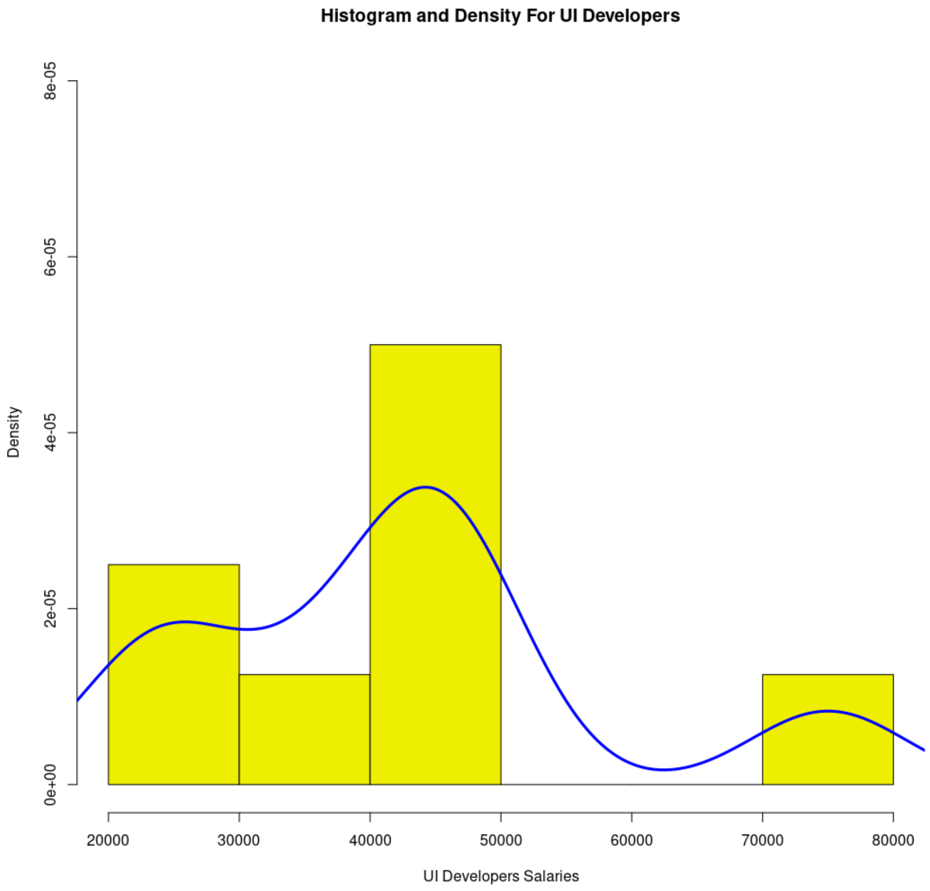
```
###############################################################################
#e
all_position_salaries = list(uid_salaries, bed_salaries, mng_salaries, hr_salaries, hse_salareis)

for (i in 1:5) {
  if(median(all_position_salaries[[i]]) > mean(all_position_salaries[[i]]))
    cat(positions[i], "salaries distribution is left-skewed", '\n')
  else
    if(median(all_position_salaries[[i]]) < mean(all_position_salaries[[i]]))
      cat(positions[i], "salaries distribution is right-skewed", '\n')
}

plot_hist_density <- function(pos, salaries) {
  hist(salaries,
       main=paste("Histogram and Density For", pos),
       xlab=paste(pos, "Salaries"),
       border="black",
       prob=TRUE,
       ylim=c(0, 0.00008),
       col="yellow2")

  lines(density(salaries), col="blue", lwd=3)
}
for (i in 1:5) {
  plot_hist_density(positions[i], all_position_salaries[[i]])
}
```

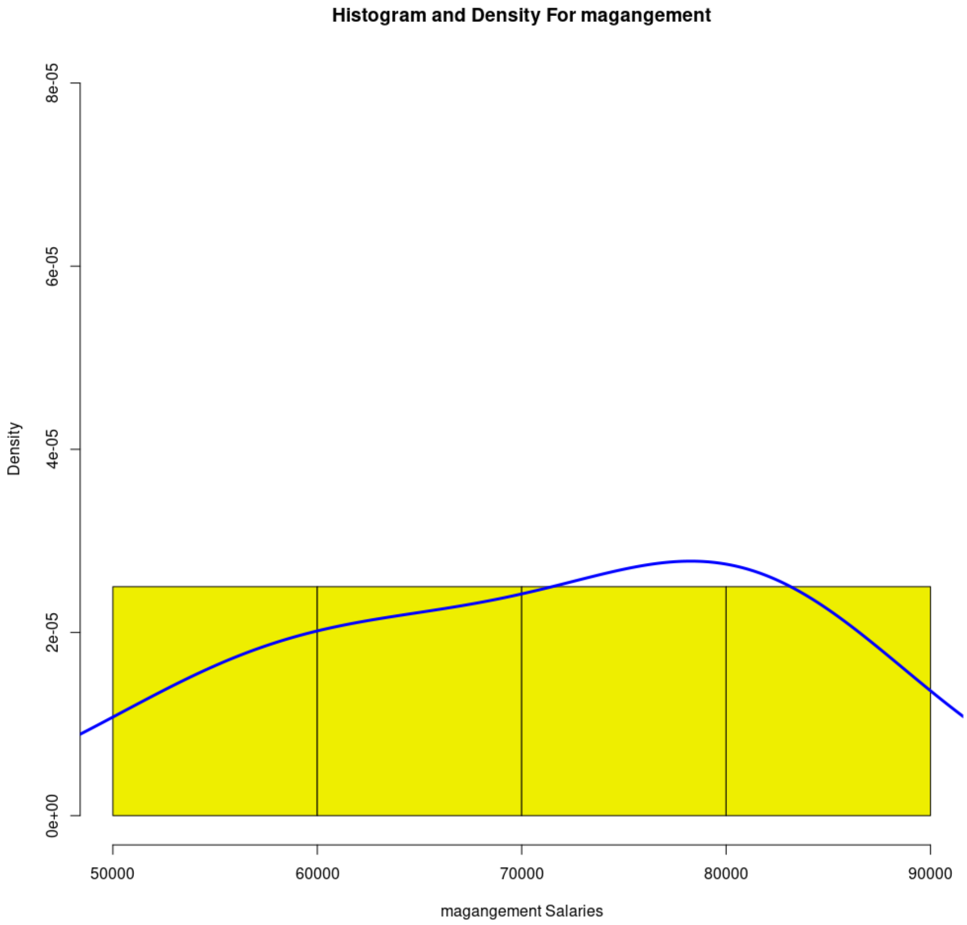
Result:
We can compare mean and median and based on the result we can decide. If median is smaller that mean → right-skewed and otherwise → left-skewed:

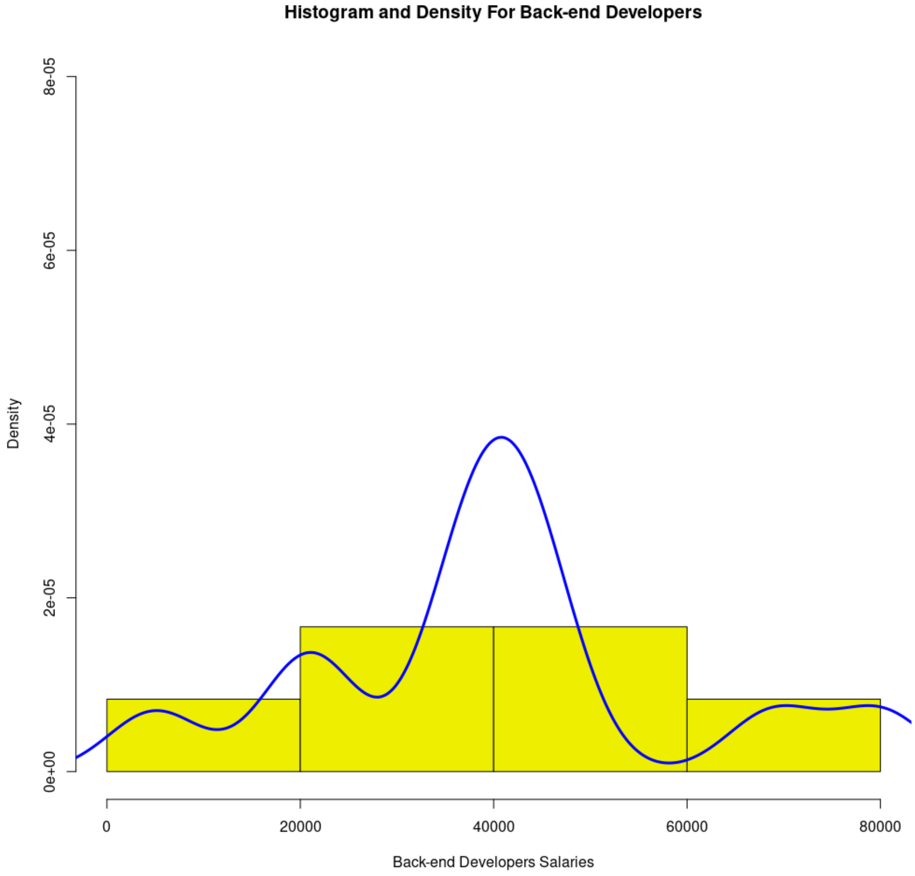


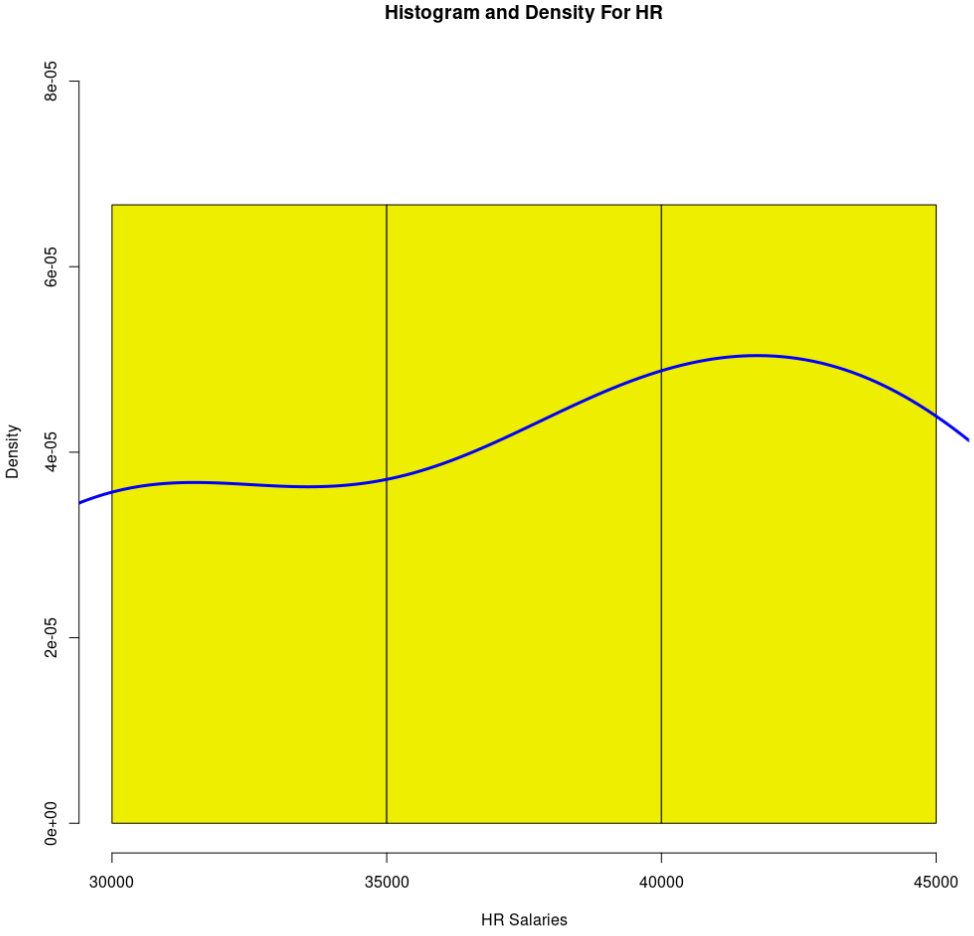
```
UI Developers salaries distribution is left-skewed
Back-end Developers salaries distribution is right-skewed
magangement salaries distribution is left-skewed
HR salaries distribution is left-skewed
HSE salaries distribution is left-skewed
```

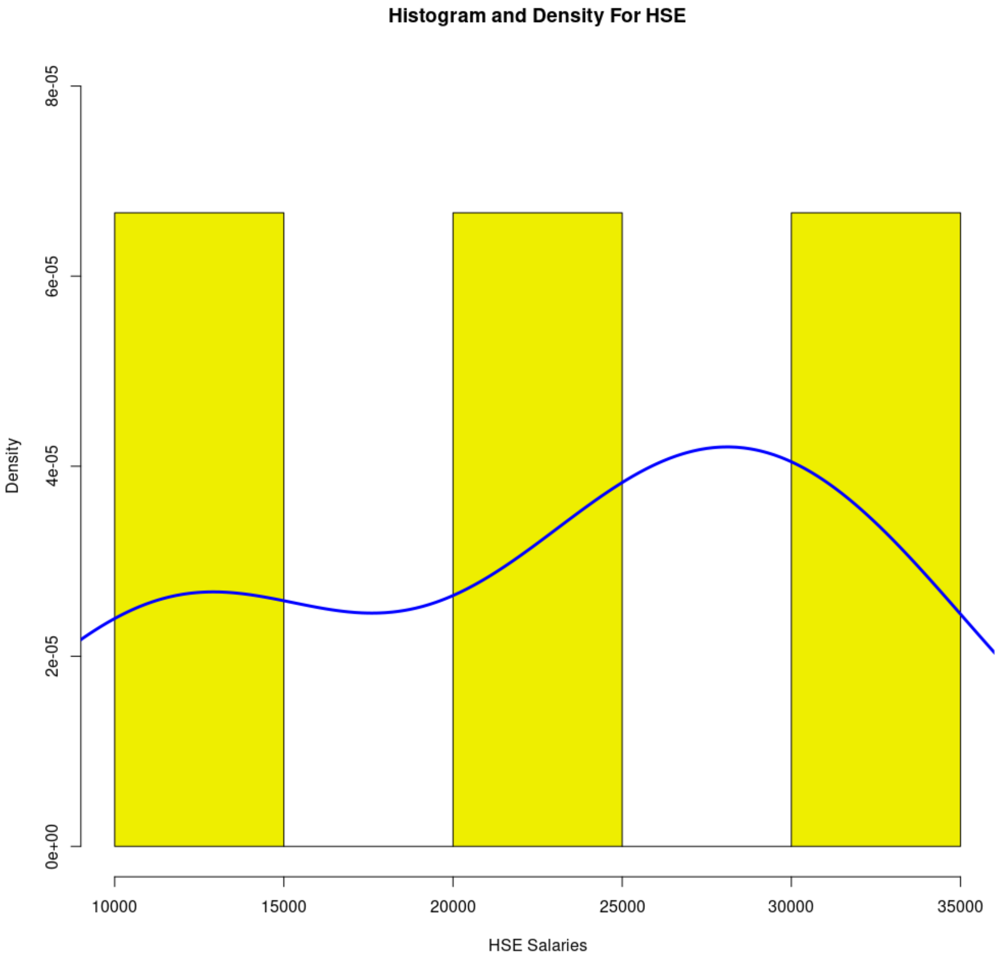
e: density graphs:



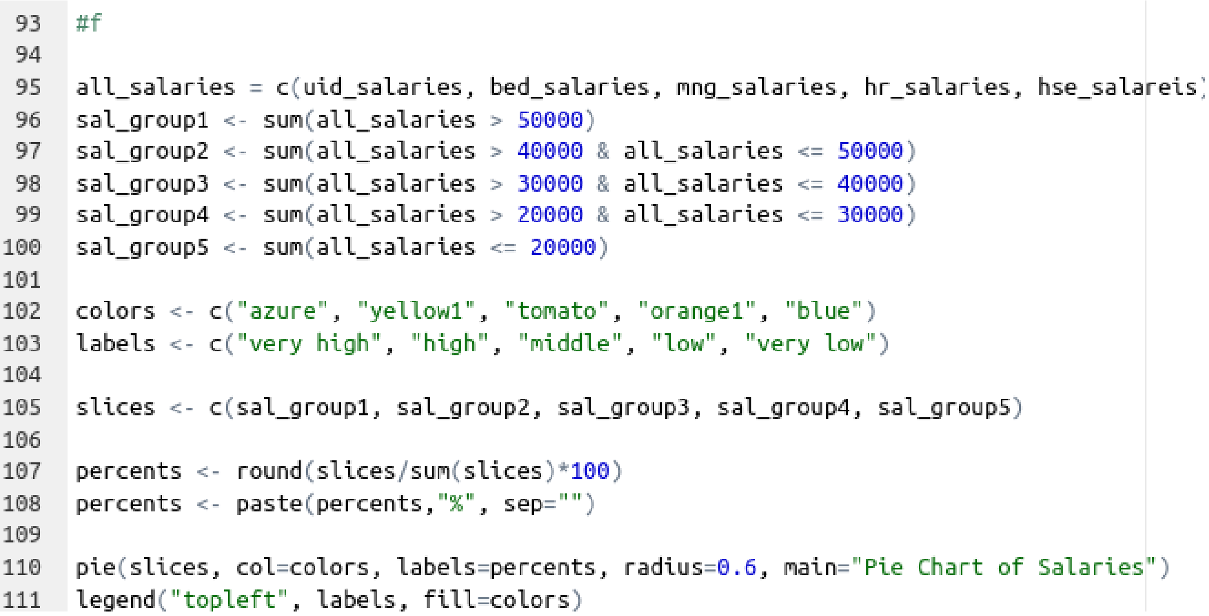



Histogram and Density For Back-end Developers
Density
0e+00
2e-05
4e-05
6e-05
8e-05
0
20000
40000
60000
80000
Back-end Developers Salaries


Histogram and Density For HR
Density
0e+00
2e-05
4e-05
6e-05
8e-05
30000
35000
40000
45000
HR Salaries

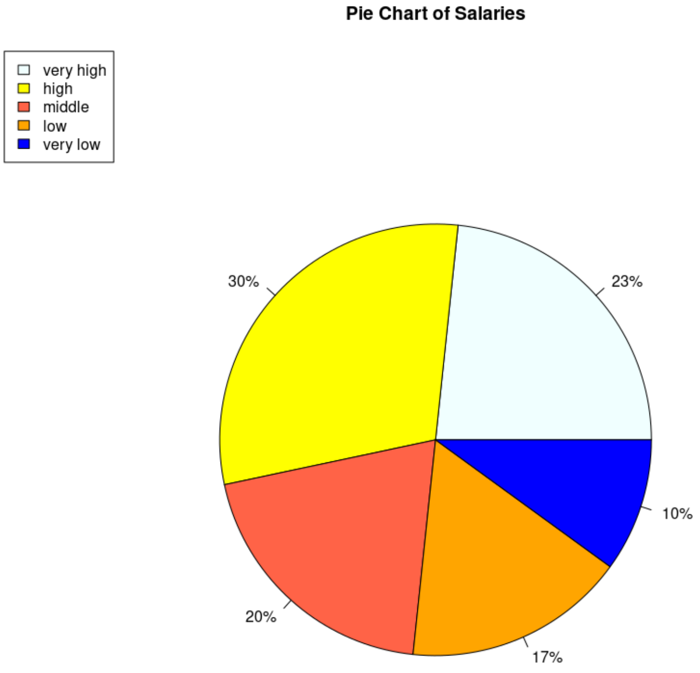

Histogram and Density For HSE
Density
0e+00
2e-05
4e-05
6e-05
8e-05
10000
15000
20000
25000
30000
35000
HSE Salaries

f.

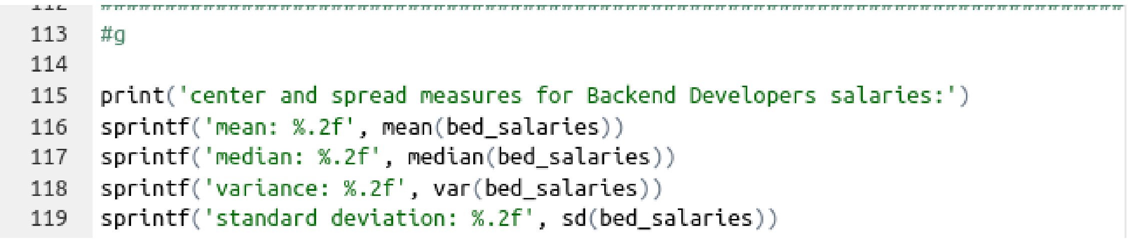


```
#f

all_salaries = c(uid_salaries, bed_salaries, mng_salaries, hr_salaries, hse_salareis)
sal_group1 <- sum(all_salaries > 50000)
sal_group2 <- sum(all_salaries > 40000 & all_salaries <= 50000)
sal_group3 <- sum(all_salaries > 30000 & all_salaries <= 40000)
sal_group4 <- sum(all_salaries > 20000 & all_salaries <= 30000)
sal_group5 <- sum(all_salaries <= 20000)

colors <- c("azure", "yellow1", "tomato", "orange1", "blue")
labels <- c("very high", "high", "middle", "low", "very low")

slices <- c(sal_group1, sal_group2, sal_group3, sal_group4, sal_group5)

percents <- round(slices/sum(slices)*100)
percents <- paste(percents,"%", sep="")

pie(slices, col=colors, labels=percents, radius=0.6, main="Pie Chart of Salaries")
legend("topleft", labels, fill=colors)
```

Result:

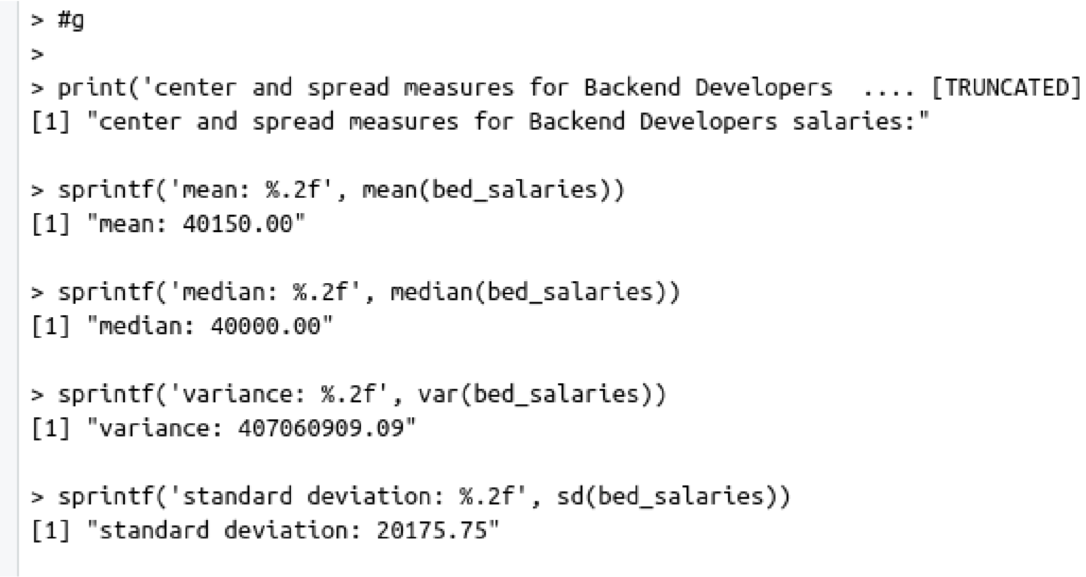


g.



```r
#g

print('center and spread measures for Backend Developers salaries:')
sprintf('mean: %.2f', mean(bed_salaries))
sprintf('median: %.2f', median(bed_salaries))
sprintf('variance: %.2f', var(bed_salaries))
sprintf('standard deviation: %.2f', sd(bed_salaries))
```

Result:



```
> #g
>
> print('center and spread measures for Backend Developers  .... [TRUNCATED]
[1] "center and spread measures for Backend Developers salaries:"

> sprintf('mean: %.2f', mean(bed_salaries))
[1] "mean: 40150.00"

> sprintf('median: %.2f', median(bed_salaries))
[1] "median: 40000.00"

> sprintf('variance: %.2f', var(bed_salaries))
[1] "variance: 407060909.09"

> sprintf('standard deviation: %.2f', sd(bed_salaries))
[1] "standard deviation: 20175.75"
```