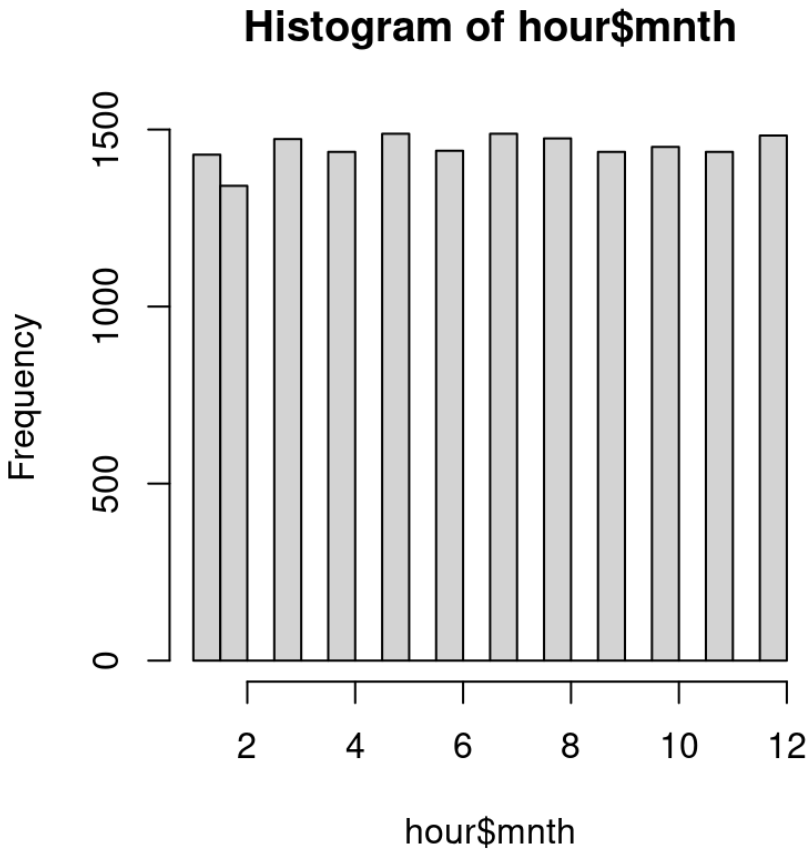
Malbora Hajdarmataj

Milestone 4

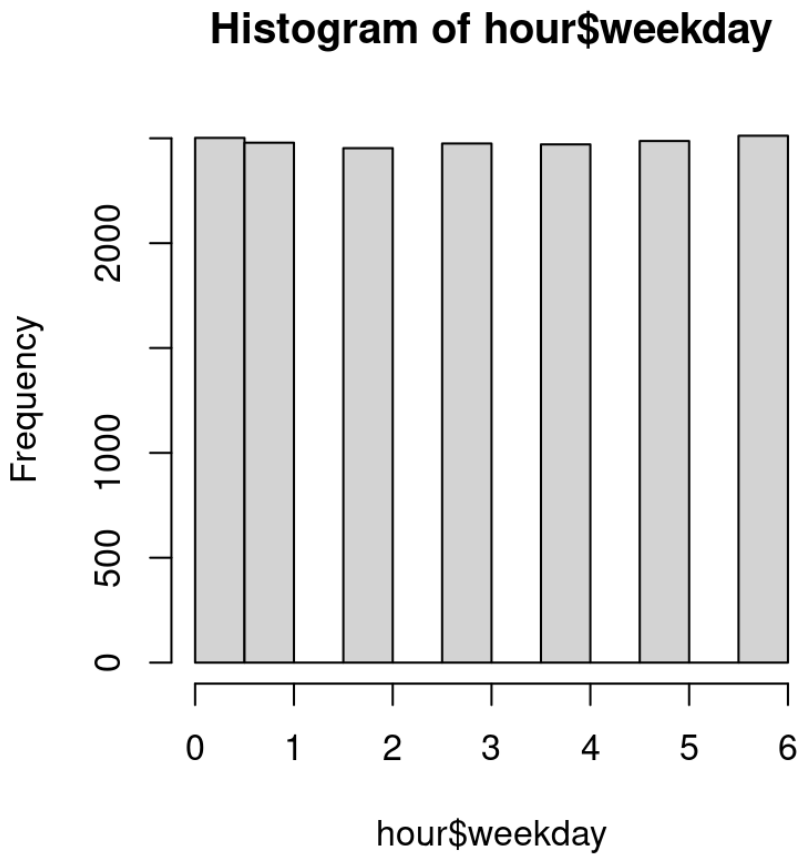
R studio link: https://rstudiow.lehman.edu/s/dbb74cfc78a317ee1c976/

For Milestone 4, I am using the same data as Milestone 3. I am using hour.cvs data and based on the distribution plots from the previous step, I can see outliers that are so extreme they could dominate the analysis. Outliner yr. and holiday has 0 observation so in this case I have missing data. I am moving the column of “yr” and “holiday” To ensure results I made the decision to remove these specific observations. This approach aligns with the recommended practices for data preprocessing, which stress the significance of handling missing data.

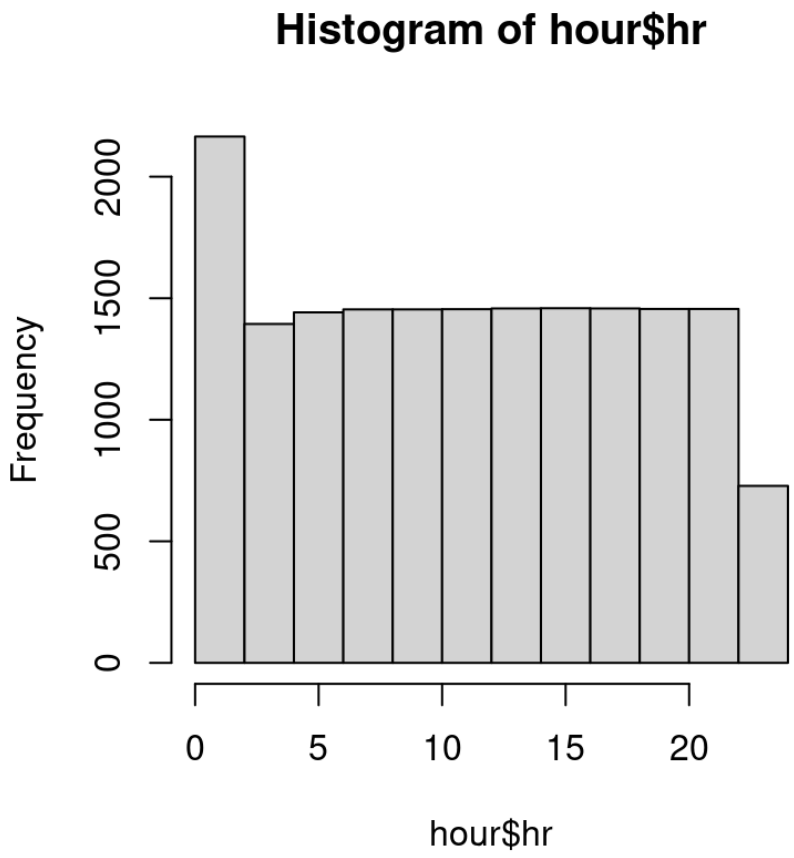
```
median(hour$mnth)
[1] 7
```



```
median(hour$weekday)
[1] 3
```




```
median(hour$weekday)
[1] 3
```



From here I can see no change in median or histogram because all the entries from yr column and holidays were 0.

	instant	dteday	season	mnth	hr	weekday	workingday	weathe…[1]	temp	atemp	hum	winds…[2]	casual
	<dbl>	<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	2011-01-01	1	1	0	6	0	1	0.24	0.288	0.81	0	3
2	2	2011-01-01	1	1	1	6	0	1	0.22	0.273	0.8	0	8
3	3	2011-01-01	1	1	2	6	0	1	0.22	0.273	0.8	0	5
4	4	2011-01-01	1	1	3	6	0	1	0.24	0.288	0.75	0	3
5	5	2011-01-01	1	1	4	6	0	1	0.24	0.288	0.75	0	0
6	6	2011-01-01	1	1	5	6	0	2	0.24	0.258	0.75	0.0896	0
7	7	2011-01-01	1	1	6	6	0	1	0.22	0.273	0.8	0	2

```
 8       8 2011-01-01      1     1     7       6          0         1  0.2  0.258  0.86  0           1
 9       9 2011-01-01      1     1     8       6          0         1  0.24 0.288  0.75  0           1
10      10 2011-01-01      1     1     9       6          0         1  0.32 0.348  0.76  0           8
# … with 17,369 more rows, 2 more variables: registered <dbl>, cnt <dbl>, and abbreviated
variab
```

New data, without two columns, yr and holiday.