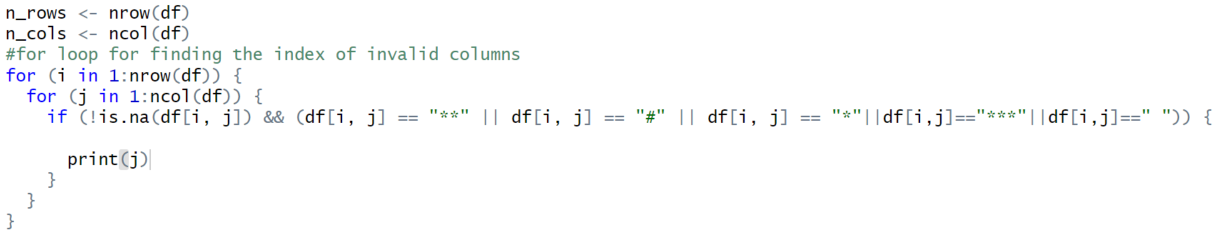
Assignment 1: R

Name: Venkata Siva Prasad kakumani

Use the files named NationalSalaries.xlsx, and Salaries.xlsx and write R scripts to perform the following tasks:

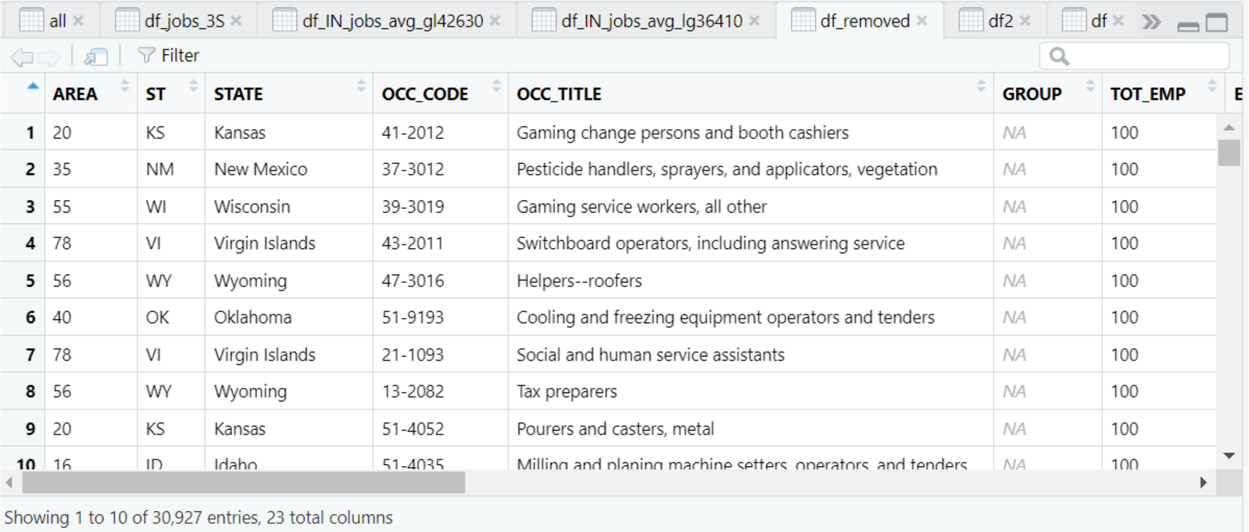
1. Data cleaning. Determine what rows have invalid entries in NationalSalaries.xlsx file and remove all such rows. (20')



```
n_rows <- nrow(df)
n_cols <- ncol(df)
#for loop for finding the index of invalid columns
for (i in 1:nrow(df)) {
  for (j in 1:ncol(df)) {
    if (!is.na(df[i, j]) && (df[i, j] == "**" || df[i, j] == "#" || df[i, j] == "*"||df[i,j]=="***"||df[i,j]==" ")) {

      print(j)
    }
  }
}
```

I got 7th column to 21st column have invalid entries. Here I excluded the NA values as some column have NA values completely. Those NA values leads to remove all the rows.

When I removed the invalid entries I got the dataset of from original dataset of 36,822 rows :

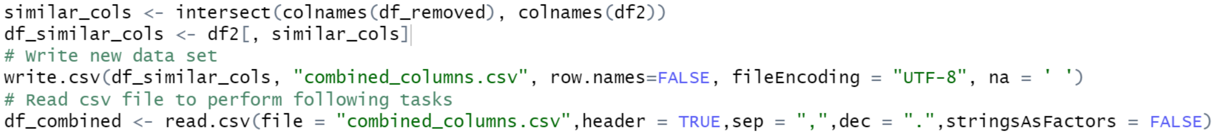


	AREA	ST	STATE	OCC_CODE	OCC_TITLE	GROUP	TOT_EMP	E
1	20	KS	Kansas	41-2012	Gaming change persons and booth cashiers	NA	100	
2	35	NM	New Mexico	37-3012	Pesticide handlers, sprayers, and applicators, vegetation	NA	100	
3	55	WI	Wisconsin	39-3019	Gaming service workers, all other	NA	100	
4	78	VI	Virgin Islands	43-2011	Switchboard operators, including answering service	NA	100	
5	56	WY	Wyoming	47-3016	Helpers--roofers	NA	100	
6	40	OK	Oklahoma	51-9193	Cooling and freezing equipment operators and tenders	NA	100	
7	78	VI	Virgin Islands	21-1093	Social and human service assistants	NA	100	
8	56	WY	Wyoming	13-2082	Tax preparers	NA	100	
9	20	KS	Kansas	51-4052	Pourers and casters, metal	NA	100	
10	16	ID	Idaho	51-4035	Milling and planing machine setters, operators, and tenders	NA	100	

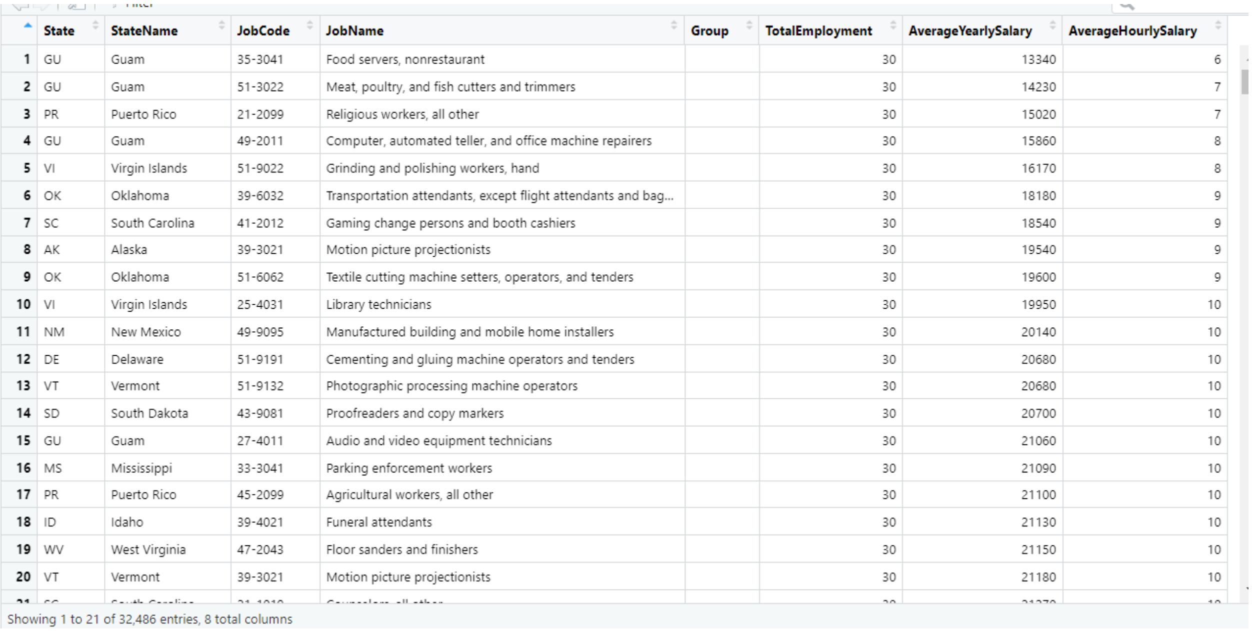
Showing 1 to 10 of 30,927 entries, 23 total columns

Total Invalid entries rows are 36822-30927=5895.

2. Select only columns that appear in the Salaries.xlsx file. Save the result into a new file and use the new file to complete the remaining tasks below.(10')



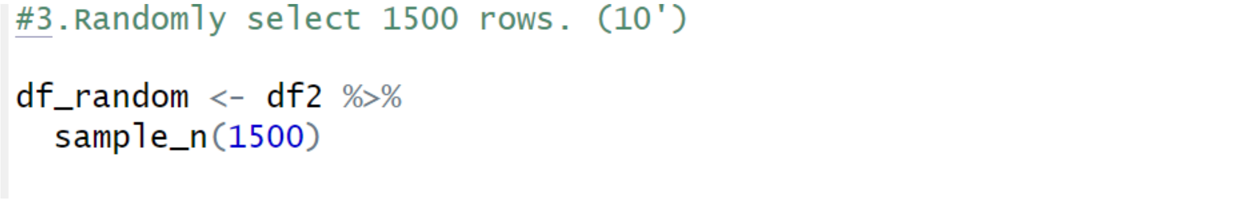
```
similar_cols <- intersect(colnames(df_removed), colnames(df2))
df_similar_cols <- df2[, similar_cols]
# Write new data set
write.csv(df_similar_cols, "combined_columns.csv", row.names=FALSE, fileEncoding = "UTF-8", na = ' ')
# Read csv file to perform following tasks
df_combined <- read.csv(file = "combined_columns.csv",header = TRUE,sep = ",",dec = ".",stringsAsFactors = FALSE)
```



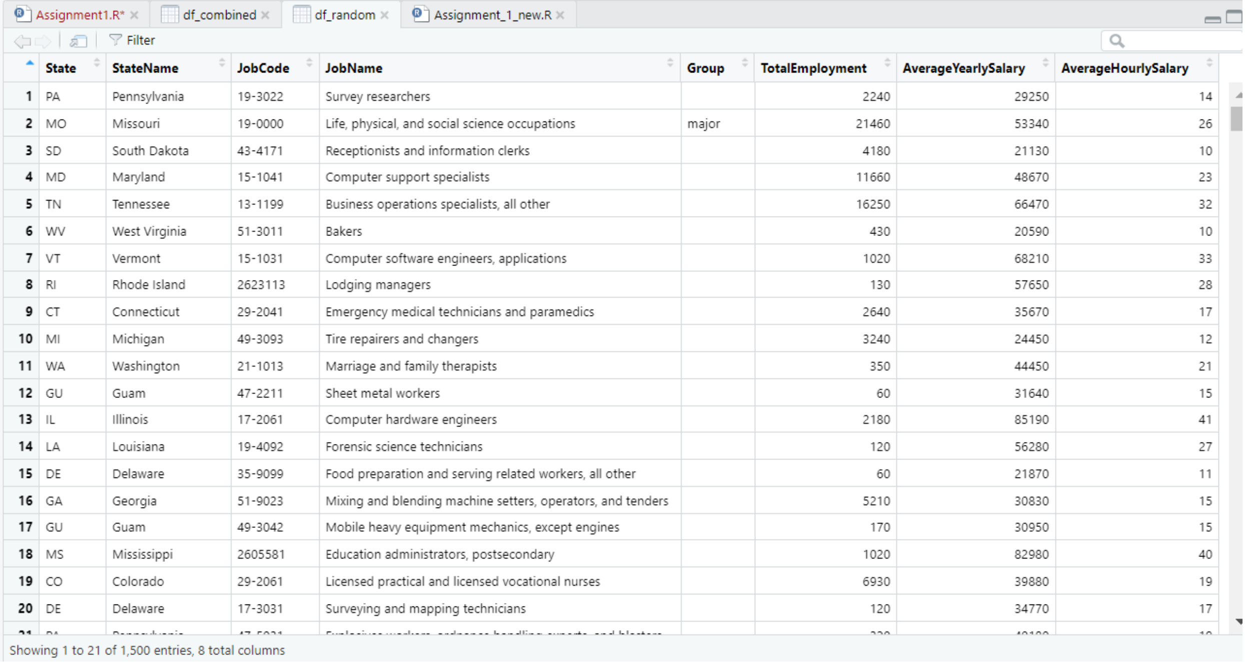
	State	StateName	JobCode	JobName	Group	TotalEmployment	AverageYearlySalary	AverageHourlySalary
1	GU	Guam	35-3041	Food servers, nonrestaurant		30	13340	6
2	GU	Guam	51-3022	Meat, poultry, and fish cutters and trimmers		30	14230	7
3	PR	Puerto Rico	21-2099	Religious workers, all other		30	15020	7
4	GU	Guam	49-2011	Computer, automated teller, and office machine repairers		30	15860	8
5	VI	Virgin Islands	51-9022	Grinding and polishing workers, hand		30	16170	8
6	OK	Oklahoma	39-6032	Transportation attendants, except flight attendants and bag...		30	18180	9
7	SC	South Carolina	41-2012	Gaming change persons and booth cashiers		30	18540	9
8	AK	Alaska	39-3021	Motion picture projectionists		30	19540	9
9	OK	Oklahoma	51-6062	Textile cutting machine setters, operators, and tenders		30	19600	9
10	VI	Virgin Islands	25-4031	Library technicians		30	19950	10
11	NM	New Mexico	49-9095	Manufactured building and mobile home installers		30	20140	10
12	DE	Delaware	51-9191	Cementing and gluing machine operators and tenders		30	20680	10
13	VT	Vermont	51-9132	Photographic processing machine operators		30	20680	10
14	SD	South Dakota	43-9081	Proofreaders and copy markers		30	20700	10
15	GU	Guam	27-4011	Audio and video equipment technicians		30	21060	10
16	MS	Mississippi	33-3041	Parking enforcement workers		30	21090	10
17	PR	Puerto Rico	45-2099	Agricultural workers, all other		30	21100	10
18	ID	Idaho	39-4021	Funeral attendants		30	21130	10
19	WV	West Virginia	47-2043	Floor sanders and finishers		30	21150	10
20	VT	Vermont	39-3021	Motion picture projectionists		30	21180	10
21	SC	South Carolina	21-1019	Counselors, all other		30	21370	10

Showing 1 to 21 of 32,486 entries, 8 total columns

3. Randomly select 1500 rows. (10')



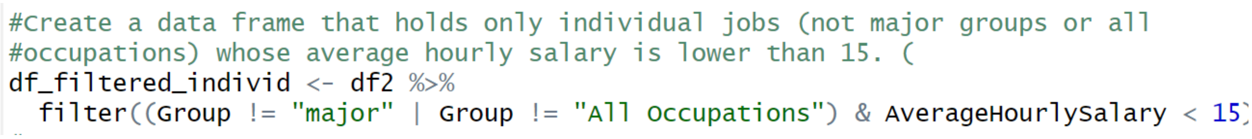
```
#3.Randomly select 1500 rows. (10')

df_random <- df2 %>%
  sample_n(1500)
```



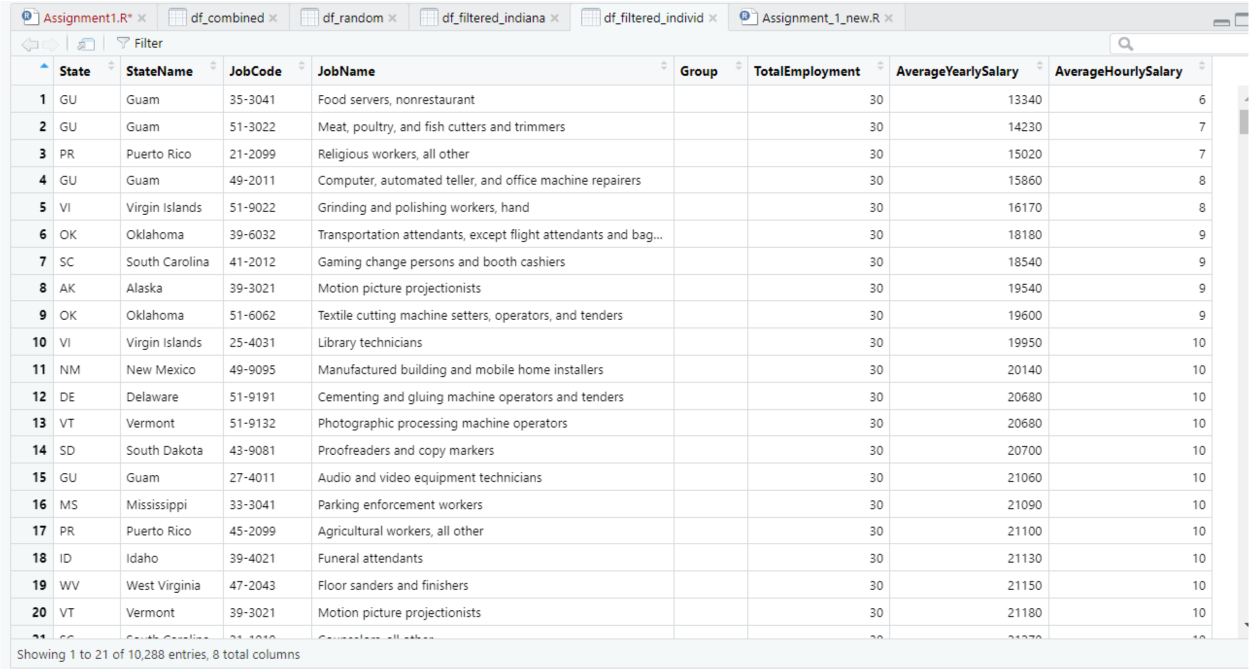
	State	StateName	JobCode	JobName	Group	TotalEmployment	AverageYearlySalary	AverageHourlySalary
1	PA	Pennsylvania	19-3022	Survey researchers		2240	29250	14
2	MO	Missouri	19-0000	Life, physical, and social science occupations	major	21460	53340	26
3	SD	South Dakota	43-4171	Receptionists and information clerks		4180	21130	10
4	MD	Maryland	15-1041	Computer support specialists		11660	48670	23
5	TN	Tennessee	13-1199	Business operations specialists, all other		16250	66470	32
6	WV	West Virginia	51-3011	Bakers		430	20590	10
7	VT	Vermont	15-1031	Computer software engineers, applications		1020	68210	33
8	RI	Rhode Island	2623113	Lodging managers		130	57650	28
9	CT	Connecticut	29-2041	Emergency medical technicians and paramedics		2640	35670	17
10	MI	Michigan	49-3093	Tire repairers and changers		3240	24450	12
11	WA	Washington	21-1013	Marriage and family therapists		350	44450	21
12	GU	Guam	47-2211	Sheet metal workers		60	31640	15
13	IL	Illinois	17-2061	Computer hardware engineers		2180	85190	41
14	LA	Louisiana	19-4092	Forensic science technicians		120	56280	27
15	DE	Delaware	35-9099	Food preparation and serving related workers, all other		60	21870	11
16	GA	Georgia	51-9023	Mixing and blending machine setters, operators, and tenders		5210	30830	15
17	GU	Guam	49-3042	Mobile heavy equipment mechanics, except engines		170	30950	15
18	MS	Mississippi	2605581	Education administrators, postsecondary		1020	82980	40
19	CO	Colorado	29-2061	Licensed practical and licensed vocational nurses		6930	39880	19
20	DE	Delaware	17-3031	Surveying and mapping technicians		120	34770	17
21	PA	Pennsylvania	47-5031	Explosives workers, ordnance handling experts, and blasters		330	40100	19

Showing 1 to 21 of 1,500 entries, 8 total columns

4. Create a data frame that holds only individual jobs (not major groups or all occupations) whose average hourly salary is lower than 15. (10')



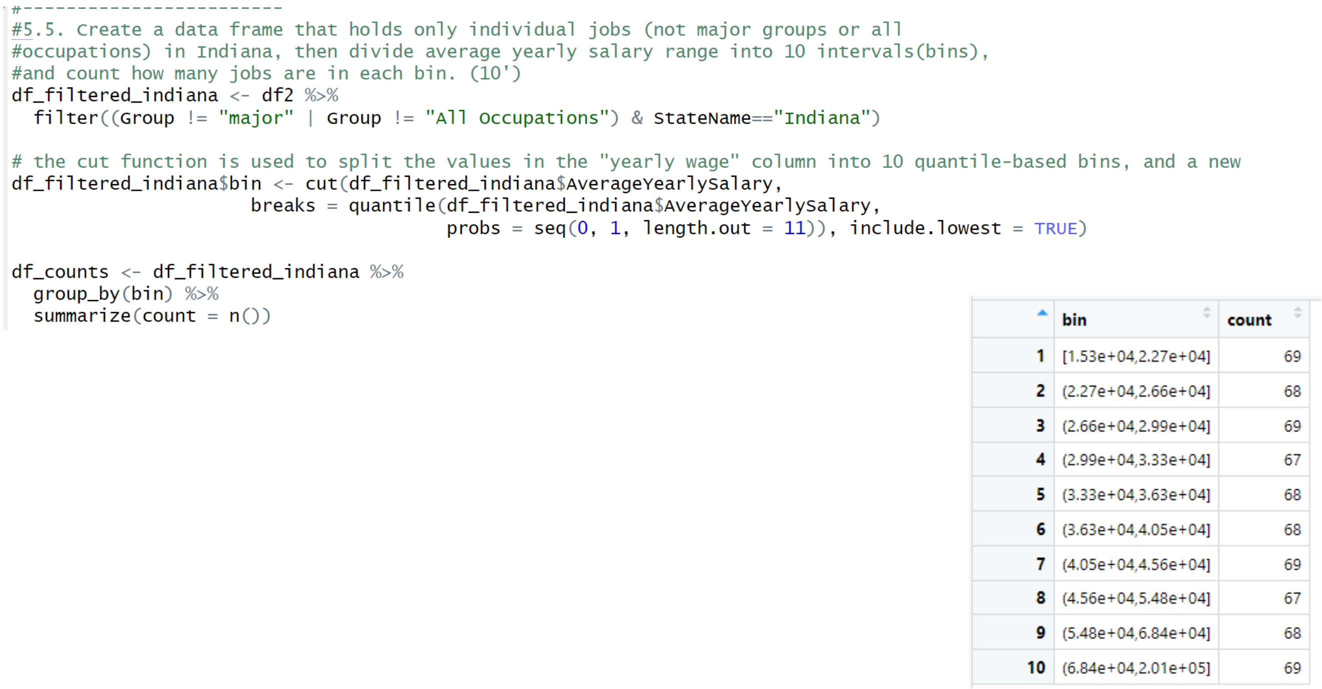
```
#Create a data frame that holds only individual jobs (not major groups or all
#occupations) whose average hourly salary is lower than 15. (
df_filtered_individ <- df2 %>%
  filter((Group != "major" | Group != "All Occupations") & AverageHourlySalary < 15
```



	State	StateName	JobCode	JobName	Group	TotalEmployment	AverageYearlySalary	AverageHourlySalary
1	GU	Guam	35-3041	Food servers, nonrestaurant		30	13340	6
2	GU	Guam	51-3022	Meat, poultry, and fish cutters and trimmers		30	14230	7
3	PR	Puerto Rico	21-2099	Religious workers, all other		30	15020	7
4	GU	Guam	49-2011	Computer, automated teller, and office machine repairers		30	15860	8
5	VI	Virgin Islands	51-9022	Grinding and polishing workers, hand		30	16170	8
6	OK	Oklahoma	39-6032	Transportation attendants, except flight attendants and bag...		30	18180	9
7	SC	South Carolina	41-2012	Gaming change persons and booth cashiers		30	18540	9
8	AK	Alaska	39-3021	Motion picture projectionists		30	19540	9
9	OK	Oklahoma	51-6062	Textile cutting machine setters, operators, and tenders		30	19600	9
10	VI	Virgin Islands	25-4031	Library technicians		30	19950	10
11	NM	New Mexico	49-9095	Manufactured building and mobile home installers		30	20140	10
12	DE	Delaware	51-9191	Cementing and gluing machine operators and tenders		30	20680	10
13	VT	Vermont	51-9132	Photographic processing machine operators		30	20680	10
14	SD	South Dakota	43-9081	Proofreaders and copy markers		30	20700	10
15	GU	Guam	27-4011	Audio and video equipment technicians		30	21060	10
16	MS	Mississippi	33-3041	Parking enforcement workers		30	21090	10
17	PR	Puerto Rico	45-2099	Agricultural workers, all other		30	21100	10
18	ID	Idaho	39-4021	Funeral attendants		30	21130	10
19	WV	West Virginia	47-2043	Floor sanders and finishers		30	21150	10
20	VT	Vermont	39-3021	Motion picture projectionists		30	21180	10

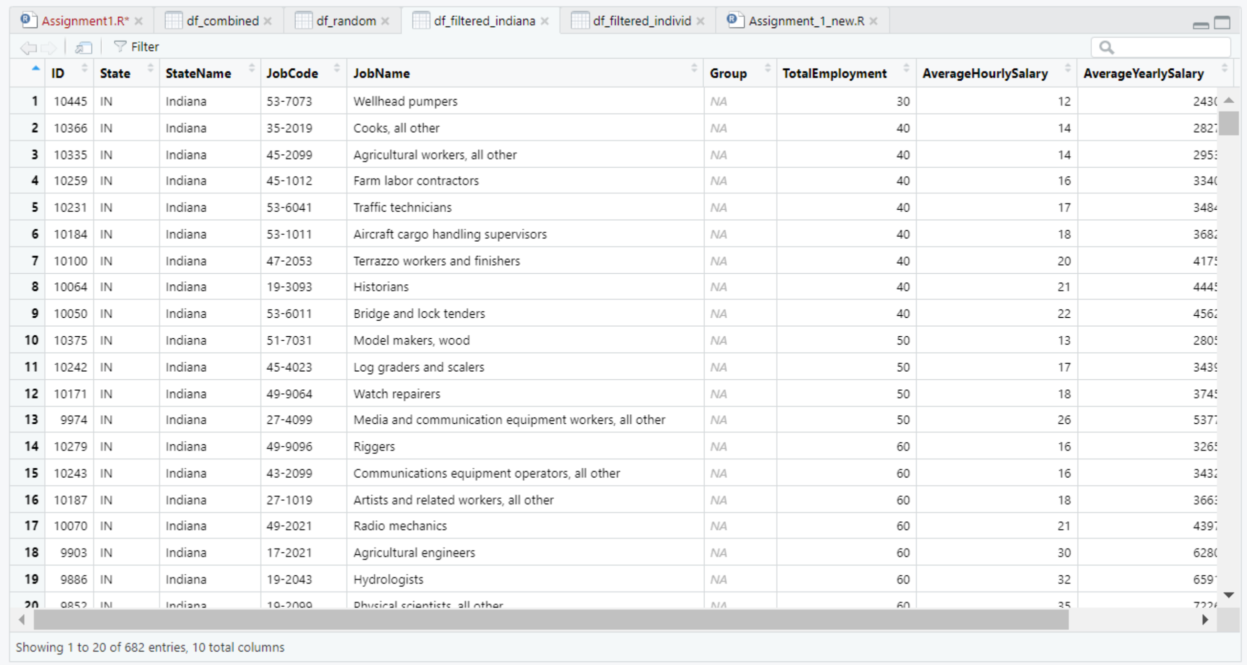
Showing 1 to 21 of 10,288 entries, 8 total columns

5. Create a data frame that holds only individual jobs (not major groups or all occupations) in Indiana, then divide average yearly salary range into 10 intervals(bins), and count how many jobs are in each bin. (10')



```
#------------------------
#5.5. Create a data frame that holds only individual jobs (not major groups or all
#occupations) in Indiana, then divide average yearly salary range into 10 intervals(bins),
#and count how many jobs are in each bin. (10')
df_filtered_indiana <- df2 %>%
  filter((Group != "major" | Group != "All Occupations") & StateName=="Indiana")

# the cut function is used to split the values in the "yearly wage" column into 10 quantile-based bins, and a new
df_filtered_indiana$bin <- cut(df_filtered_indiana$AverageYearlySalary,
                      breaks = quantile(df_filtered_indiana$AverageYearlySalary,
                                        probs = seq(0, 1, length.out = 11)), include.lowest = TRUE)

df_counts <- df_filtered_indiana %>%
  group_by(bin) %>%
  summarize(count = n())
```

	bin	count
1	[1.53e+04,2.27e+04]	69
2	(2.27e+04,2.66e+04]	68
3	(2.66e+04,2.99e+04]	69
4	(2.99e+04,3.33e+04]	67
5	(3.33e+04,3.63e+04]	68
6	(3.63e+04,4.05e+04]	68
7	(4.05e+04,4.56e+04]	69
8	(4.56e+04,5.48e+04]	67
9	(5.48e+04,6.84e+04]	68
10	(6.84e+04,2.01e+05]	69

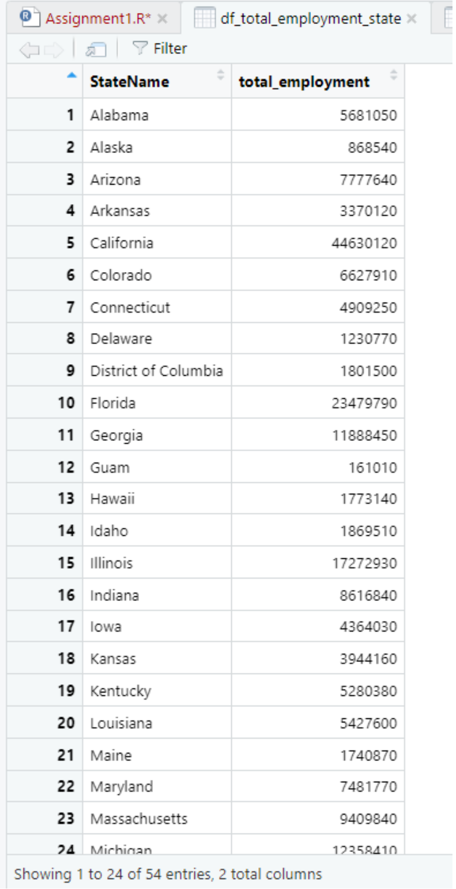


	ID	State	StateName	JobCode	JobName	Group	TotalEmployment	AverageHourlySalary	AverageYearlySalary
1	10445	IN	Indiana	53-7073	Wellhead pumpers	NA	30	12	243
2	10366	IN	Indiana	35-2019	Cooks, all other	NA	40	14	282
3	10335	IN	Indiana	45-2099	Agricultural workers, all other	NA	40	14	295
4	10259	IN	Indiana	45-1012	Farm labor contractors	NA	40	16	334
5	10231	IN	Indiana	53-6041	Traffic technicians	NA	40	17	348
6	10184	IN	Indiana	53-1011	Aircraft cargo handling supervisors	NA	40	18	368
7	10100	IN	Indiana	47-2053	Terrazzo workers and finishers	NA	40	20	417
8	10064	IN	Indiana	19-3093	Historians	NA	40	21	444
9	10050	IN	Indiana	53-6011	Bridge and lock tenders	NA	40	22	456
10	10375	IN	Indiana	51-7031	Model makers, wood	NA	50	13	280
11	10242	IN	Indiana	45-4023	Log graders and scalers	NA	50	17	343
12	10171	IN	Indiana	49-9064	Watch repairers	NA	50	18	374
13	9974	IN	Indiana	27-4099	Media and communication equipment workers, all other	NA	50	26	537
14	10279	IN	Indiana	49-9096	Riggers	NA	60	16	326
15	10243	IN	Indiana	43-2099	Communications equipment operators, all other	NA	60	16	343
16	10187	IN	Indiana	27-1019	Artists and related workers, all other	NA	60	18	366
17	10070	IN	Indiana	49-2021	Radio mechanics	NA	60	21	439
18	9903	IN	Indiana	17-2021	Agricultural engineers	NA	60	30	628
19	9886	IN	Indiana	19-2043	Hydrologists	NA	60	32	659
20	9852	IN	Indiana	19-2099	Physical scientists, all other	NA	60	35	722

Showing 1 to 20 of 682 entries, 10 total columns

6. Find the total employment for each state. (10')

```
#6. Find the total employment for each state.
df_total_employment_state <- df2 %>%
  group_by(StateName) %>%
  summarize(total_employment = sum(TotalEmployment))
```

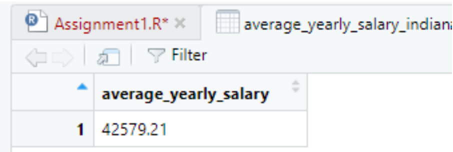


	StateName	total_employment
1	Alabama	5681050
2	Alaska	868540
3	Arizona	7777640
4	Arkansas	3370120
5	California	44630120
6	Colorado	6627910
7	Connecticut	4909250
8	Delaware	1230770
9	District of Columbia	1801500
10	Florida	23479790
11	Georgia	11888450
12	Guam	161010
13	Hawaii	1773140
14	Idaho	1869510
15	Illinois	17272930
16	Indiana	8616840
17	Iowa	4364030
18	Kansas	3944160
19	Kentucky	5280380
20	Louisiana	5427600
21	Maine	1740870
22	Maryland	7481770
23	Massachusetts	9409840
24	Michigan	12358410

Showing 1 to 24 of 54 entries, 2 total columns

7. Find the average yearly salary of all jobs in Indiana, and compare it with data provided in the data set (42630 vs 36410). (20’)

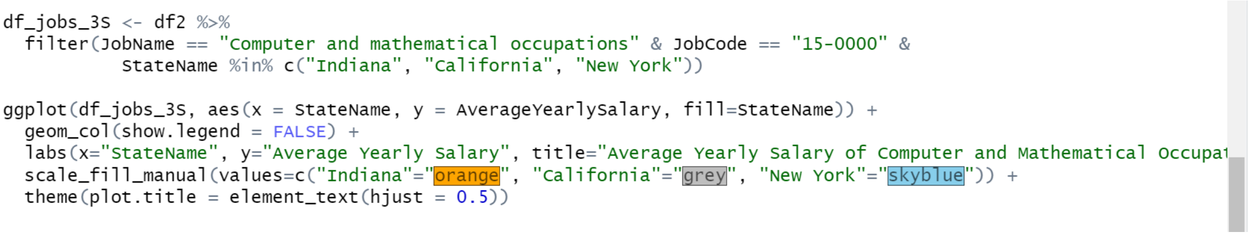
```r
#7. Find the average yearly salary of all jobs in Indiana, and compare it with data
#provided in the data set (42630 vs 36410). (20’)
df_indiana_jobs <- df2 %>%
  filter(StateName == "Indiana")

average_yearly_salary_indiana <- df_indiana_jobs %>%
  summarize(average_yearly_salary = mean(AverageYearlySalary))

#comparing the average salary with 42630 and 36410
df_IN_jobs_g <- df_indiana_jobs %>%
  filter(AverageYearlySalary > 42630)

df_IN_jobs_l <- df_indiana_jobs %>%
  filter(AverageYearlySalary < 42630)

df_IN_jobs_l3 <- df_indiana_jobs %>%
  filter(AverageYearlySalary < 36410)

df_IN_jobs_g3 <- df_indiana_jobs %>%
  filter(AverageYearlySalary > 36410)

df_IN_jobs_gl <- df_indiana_jobs %>%
  filter(AverageYearlySalary > 36410 & AverageYearlySalary < 42630)
df_IN_jobs_avg_lg36410 <- df_indiana_jobs %>%
  filter(AverageYearlySalary > 36410 & AverageYearlySalary < mean(AverageYearlySalary))
```



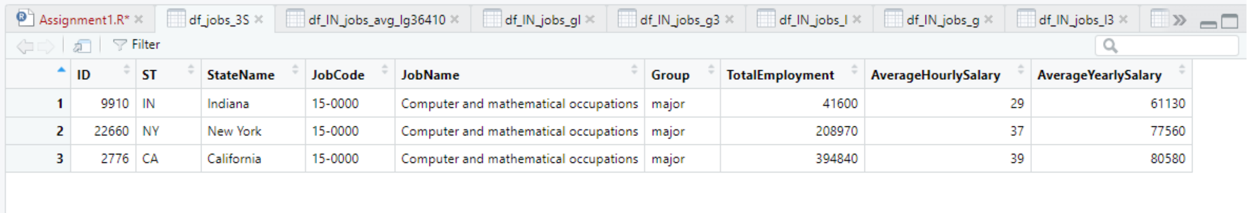
	average_yearly_salary
1	42579.21

On comparison with 42630 and 36410 values, the salaries greater than 42630 are 245 and less than are 438. In the same way, for 36410 are 340 and 341, respectively. 95 rows are there in between them. Compared to average yearly salary of Indiana, 95 rows are there in between average of Indiana and 36410. 0 rows are there in between average of Indiana and 42630.

8. Use a chart to compare average yearly salaries of "Computer and mathematical occupations" (coded 15 - xxxx) in Indiana, California and New York. Use colors and legends to make your chart informative. (10’)

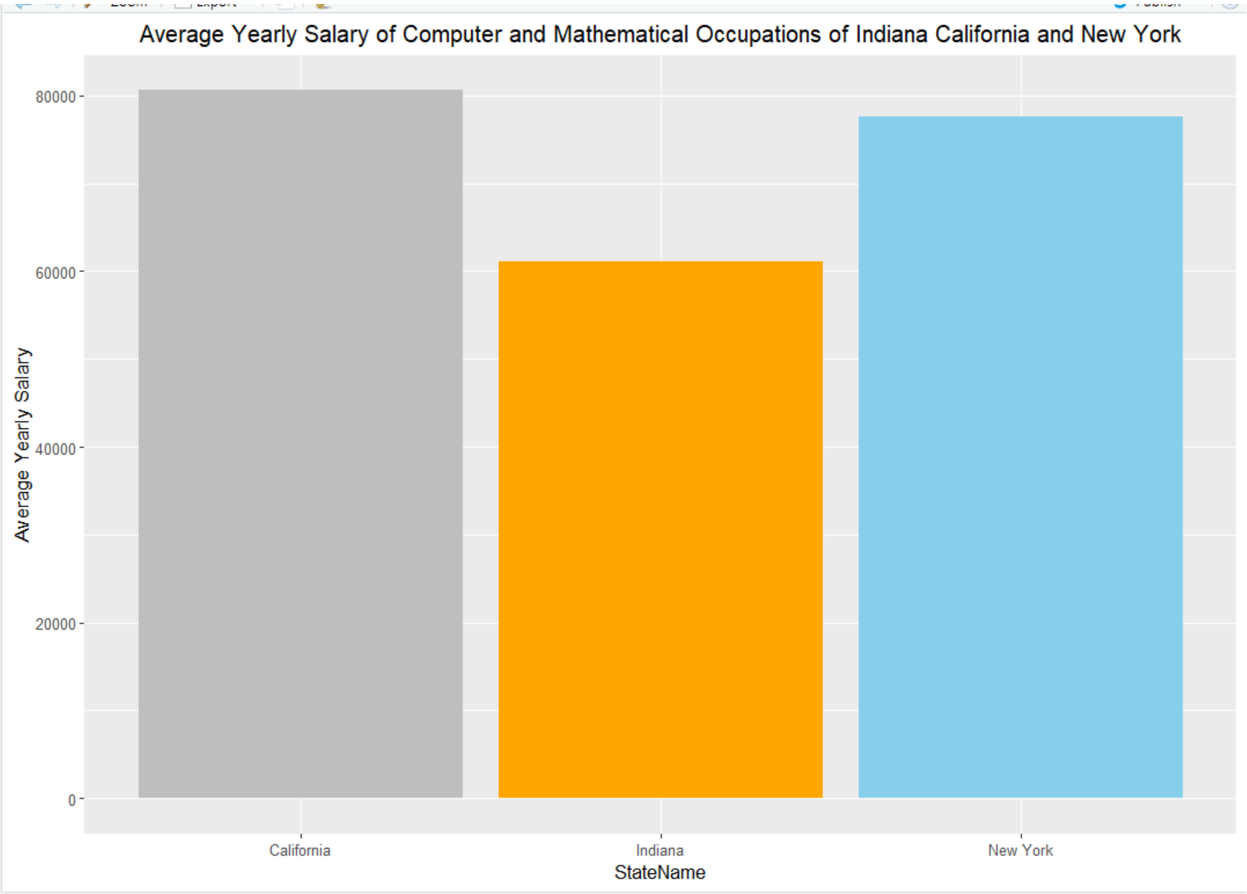


```r
df_jobs_3S <- df2 %>%
  filter(JobName == "Computer and mathematical occupations" & JobCode == "15-0000" &
           StateName %in% c("Indiana", "California", "New York"))

ggplot(df_jobs_3S, aes(x = StateName, y = AverageYearlySalary, fill=StateName)) +
  geom_col(show.legend = FALSE) +
  labs(x="StateName", y="Average Yearly Salary", title="Average Yearly Salary of Computer and Mathematical Occupat
  scale_fill_manual(values=c("Indiana"="orange", "California"="grey", "New York"="skyblue")) +
  theme(plot.title = element_text(hjust = 0.5))
```



	ID	ST	StateName	JobCode	JobName	Group	TotalEmployment	AverageHourlySalary	AverageYearlySalary
1	9910	IN	Indiana	15-0000	Computer and mathematical occupations	major	41600	29	61130
2	22660	NY	New York	15-0000	Computer and mathematical occupations	major	208970	37	77560
3	2776	CA	California	15-0000	Computer and mathematical occupations	major	394840	39	80580

Plot:


Average Yearly Salary of Computer and Mathematical Occupations of Indiana California and New York
Average Yearly Salary
California
Indiana
New York
StateName