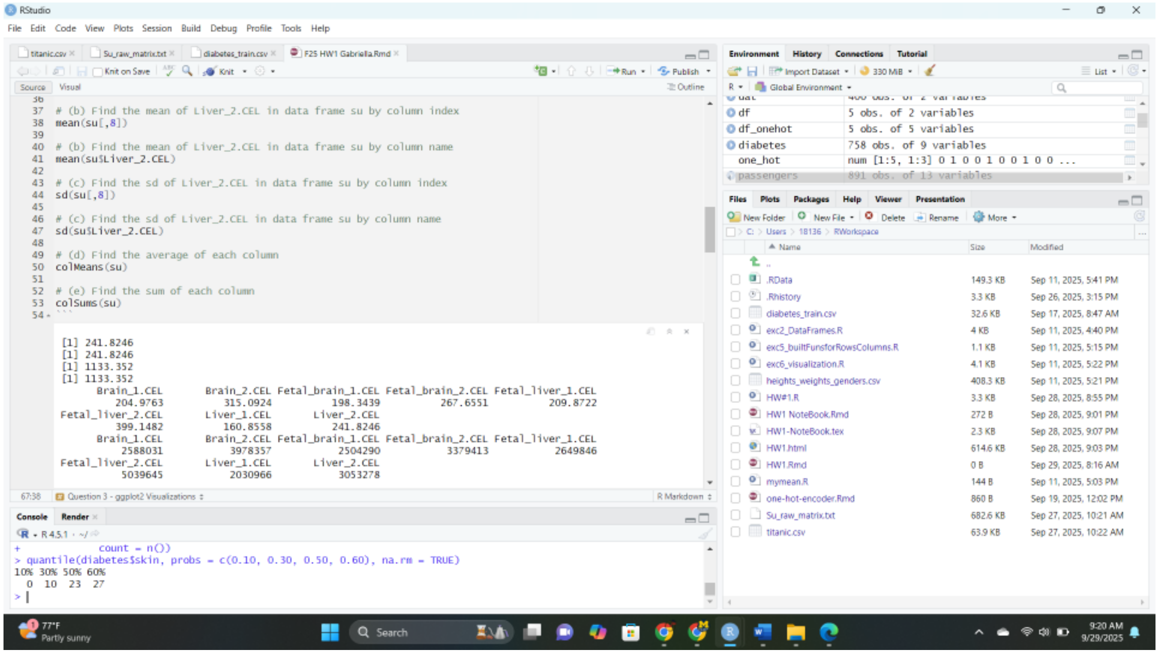
Explanation of R Commands in Notebook (HW1)

General Setup

- `knitr::opts_chunk$set(echo = TRUE)` → tells knitr to display code along with results.
- `library(ggplot2)` → loads the ggplot2 package (graphics/plotting).
- `library(tinytex)` → loads tinytex, used for LaTeX/PDF rendering.
- `library(dplyr)` → loads dplyr for data manipulation.
- `library(tidyverse)` → loads the tidyverse collection of packages.
- `set.seed(123)` → makes random number generation reproducible.

Question 1 – Su_raw_matrix.txt

- `read.delim()` → reads a tab-delimited file into a data frame.
- `mean(su[,8])` → mean of the 8th column.
- `mean(su$Liver_2.CEL)` → mean of column by name.
- `sd()` → standard deviation of a vector.
- `colMeans(su)` → column-wise means.
- `colSums(su)` → column-wise sums.



```
# (b) Find the mean of Liver_2.CEL in data frame su by column index
mean(su[,8])

# (b) Find the mean of Liver_2.CEL in data frame su by column name
mean(su$Liver_2.CEL)

# (c) Find the sd of Liver_2.CEL in data frame su by column index
sd(su[,8])

# (c) Find the sd of Liver_2.CEL in data frame su by column name
sd(su$Liver_2.CEL)

# (d) Find the average of each column
colMeans(su)

# (e) Find the sum of each column
colSums(su)
```
```

[1] 241.8246
[1] 241.8246
[1] 1133.352
[1] 1133.352
      Brain_1.CEL       Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL Fetal_liver_1.CEL
         204.9763          315.0924          198.3439          267.6551          209.8722
Fetal_liver_2.CEL       Liver_1.CEL       Liver_2.CEL
         399.1482          160.8558          241.8246
      Brain_1.CEL       Brain_2.CEL Fetal_brain_1.CEL Fetal_brain_2.CEL Fetal_liver_1.CEL
          2588031           3978357           2504290           3379413           2649846
Fetal_liver_2.CEL       Liver_1.CEL       Liver_2.CEL
          5039645           2030966           3053278
```

```
+           count = n())
> quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
10% 30% 50% 60% 
  0  10  23  27 
> |
```

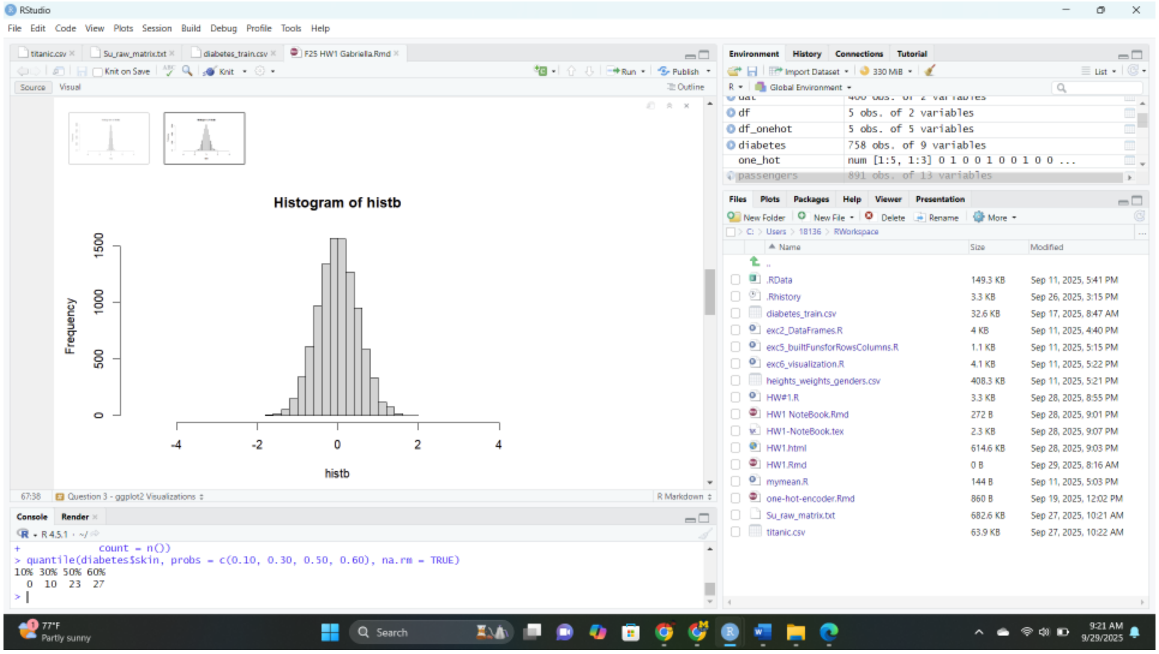
df	5 obs. of 2 variables
df_onehot	5 obs. of 5 variables
diabetes	758 obs. of 9 variables
one_hot	num [1:5, 1:3] 0 1 0 0 1 0 0 1 0 0 ...
passengers	891 obs. of 13 variables

Name	Size	Modified
..		
.RData	149.3 KB	Sep 11, 2025, 5:41 PM
.Rhistory	3.3 KB	Sep 26, 2025, 3:15 PM
diabetes_train.csv	32.6 KB	Sep 17, 2025, 8:47 AM
exc2_DataFrames.R	4 KB	Sep 11, 2025, 4:40 PM
exc5_builtFunsforRowsColumns.R	1.1 KB	Sep 11, 2025, 5:15 PM
exc6_visualization.R	4.1 KB	Sep 11, 2025, 5:22 PM
heights_weights_genders.csv	408.3 KB	Sep 11, 2025, 5:21 PM
HW#1.R	3.3 KB	Sep 28, 2025, 8:55 PM
HW1 NoteBook.Rmd	272 B	Sep 28, 2025, 9:01 PM
HW1-NoteBook.tex	2.3 KB	Sep 28, 2025, 9:07 PM
HW1.html	614.6 KB	Sep 28, 2025, 9:03 PM
HW1.Rmd	0 B	Sep 29, 2025, 8:16 AM
mymean.R	144 B	Sep 11, 2025, 5:03 PM
one-hot-encoder.Rmd	860 B	Sep 19, 2025, 12:02 PM
Su_raw_matrix.txt	682.6 KB	Sep 27, 2025, 10:21 AM
titanic.csv	63.9 KB	Sep 27, 2025, 10:22 AM

Question 2 – Random Numbers & Histograms

- `rnorm(n, mean, sd)` → generates normal random numbers.

- `hist()` → plots histogram.
- `xlim = c(-5,5)` → sets x-axis limits.



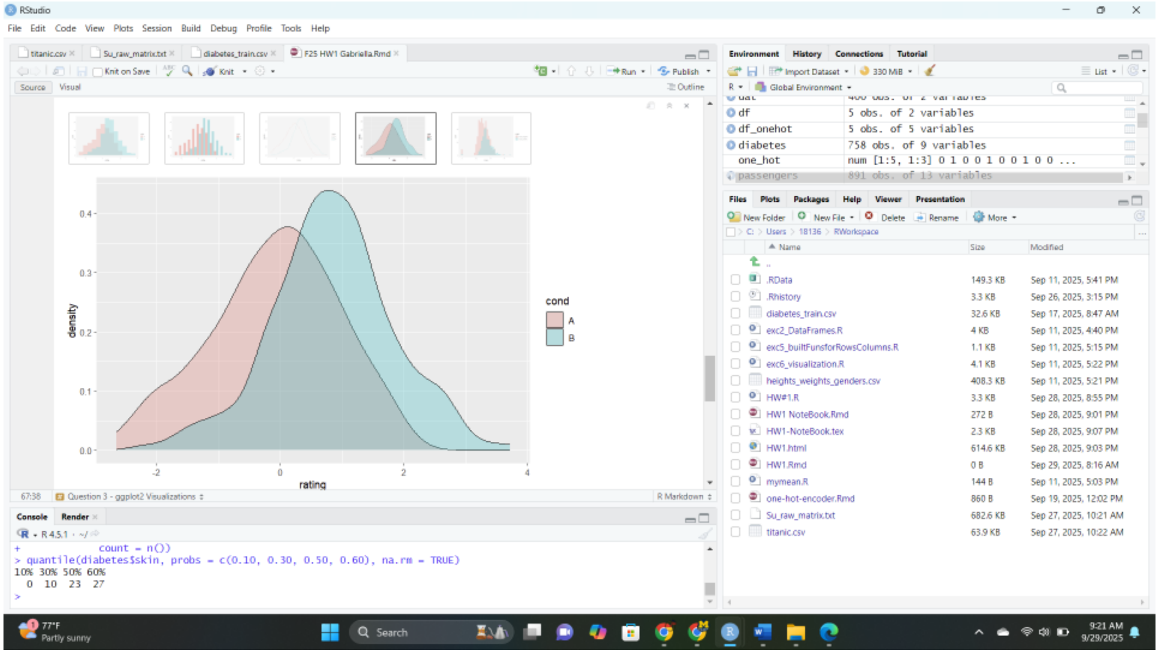
```
+               count = n())
> quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
10% 30% 50% 60%
  0  10  23  27
> |
```

dat	400 obs. of 2 variables
df	5 obs. of 2 variables
df_onehot	5 obs. of 5 variables
diabetes	758 obs. of 9 variables
one_hot	num [1:5, 1:3] 0 1 0 0 1 0 0 1 0 0 ...
passengers	891 obs. of 13 variables

Name	Size	Modified
..		
.RData	149.3 KB	Sep 11, 2025, 5:41 PM
.Rhistory	3.3 KB	Sep 26, 2025, 3:15 PM
diabetes_train.csv	32.6 KB	Sep 17, 2025, 8:47 AM
exc2_DataFrames.R	4 KB	Sep 11, 2025, 4:40 PM
exc5_builtFunsforRowsColumns.R	1.1 KB	Sep 11, 2025, 5:15 PM
exc6_visualization.R	4.1 KB	Sep 11, 2025, 5:22 PM
heights_weights_genders.csv	408.3 KB	Sep 11, 2025, 5:21 PM
HW#1.R	3.3 KB	Sep 28, 2025, 8:55 PM
HW1 NoteBook.Rmd	272 B	Sep 28, 2025, 9:01 PM
HW1-NoteBook.tex	2.3 KB	Sep 28, 2025, 9:07 PM
HW1.html	614.6 KB	Sep 28, 2025, 9:03 PM
HW1.Rmd	0 B	Sep 29, 2025, 8:16 AM
mymean.R	144 B	Sep 11, 2025, 5:03 PM
one-hot-encoder.Rmd	860 B	Sep 19, 2025, 12:02 PM
Su_raw_matrix.txt	682.6 KB	Sep 27, 2025, 10:21 AM
titanic.csv	63.9 KB	Sep 27, 2025, 10:22 AM

Question 3 – ggplot2 Visualizations

- `data.frame()` → creates a data frame.
- `rep()` → repeats values.
- `factor()` → categorical variable.
- `ggplot()` → initializes plot.
- `aes()` → sets aesthetics mapping.
- `geom_histogram()` → histogram layer.
- `geom_density()` → smooth density plot.
- `alpha` → transparency level.
- `position = "identity"` → overlay plots.
- `position = "dodge"` → side-by-side plots.
- `read.csv()` → loads CSV file.
- `as.factor()` → converts variable to factor.



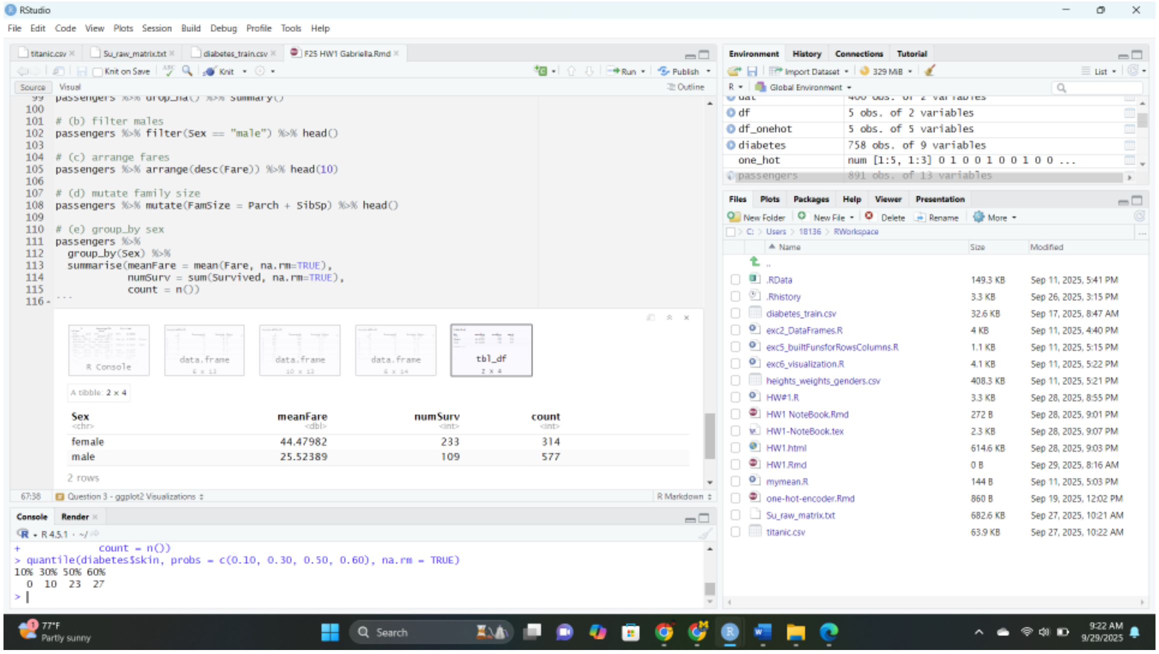
dat	400 obs. of 2 variables
df	5 obs. of 2 variables
df_onehot	5 obs. of 5 variables
diabetes	758 obs. of 9 variables
one_hot	num [1:5, 1:3] 0 1 0 0 1 0 0 1 0 0 ...
passengers	891 obs. of 13 variables

Name	Size	Modified
..		
.RData	149.3 KB	Sep 11, 2025, 5:41 PM
.Rhistory	3.3 KB	Sep 26, 2025, 3:15 PM
diabetes_train.csv	32.6 KB	Sep 17, 2025, 8:47 AM
exc2_DataFrames.R	4 KB	Sep 11, 2025, 4:40 PM
exc5_builtFunsforRowsColumns.R	1.1 KB	Sep 11, 2025, 5:15 PM
exc6_visualization.R	4.1 KB	Sep 11, 2025, 5:22 PM
heights_weights_genders.csv	408.3 KB	Sep 11, 2025, 5:21 PM
HW#1.R	3.3 KB	Sep 28, 2025, 8:55 PM
HW1 NoteBook.Rmd	272 B	Sep 28, 2025, 9:01 PM
HW1-NoteBook.tex	2.3 KB	Sep 28, 2025, 9:07 PM
HW1.html	614.6 KB	Sep 28, 2025, 9:03 PM
HW1.Rmd	0 B	Sep 29, 2025, 8:16 AM
mymean.R	144 B	Sep 11, 2025, 5:03 PM
one-hot-encoder.Rmd	860 B	Sep 19, 2025, 12:02 PM
Su_raw_matrix.txt	682.6 KB	Sep 27, 2025, 10:21 AM
titanic.csv	63.9 KB	Sep 27, 2025, 10:22 AM

```
+             count = n())
> quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
10% 30% 50% 60% 
  0  10  23  27 
> 
```

Question 4 – Titanic Data with dplyr

- `%>%` → pipe operator (passes result).
- `drop_na()` → removes rows with missing values.
- `summary()` → summary stats.
- `filter()` → keeps rows meeting condition.
- `head()` → first few rows.
- `arrange(desc())` → sort descending.
- `mutate()` → add new column.
- `group_by()` → group data by variable.
- `summarise()` → compute summary statistics:
  - `mean()` → mean of variable.
  - `sum()` → sum of variable.
  - `n()` → count rows.



```r
passengers %>% drop_na() %>% summary()

# (b) filter males
passengers %>% filter(Sex == "male") %>% head()

# (c) arrange fares
passengers %>% arrange(desc(Fare)) %>% head(10)

# (d) mutate family size
passengers %>% mutate(FamSize = Parch + SibSp) %>% head()

# (e) group_by sex
passengers %>%
  group_by(Sex) %>%
  summarise(meanFare = mean(Fare, na.rm=TRUE),
            numSurv = sum(Survived, na.rm=TRUE),
            count = n())
```
```

R Console
data.frame
6 x 13
data.frame
10 x 13
data.frame
6 x 14
tbl_df
2 x 4

A tibble: 2 x 4

| Sex<br>&lt;chr&gt; | meanFare<br>&lt;dbl&gt; | numSurv<br>&lt;int&gt; | count<br>&lt;int&gt; |
|---|---|---|---|
| female | 44.47982 | 233 | 314 |
| male | 25.52389 | 109 | 577 |

2 rows

67:38
Question 3 - ggplot2 Visualizations
R Markdown
Console
Render
R 4.5.1 · ~/

```
+   count = n())
> quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
10% 30% 50% 60%
  0 10 23 27
>
~~~

Environment
History
Connections
Tutorial
Import Dataset
329 MiB
List
R
Global Environment

dat	400 obs. of 2 variables
df	5 obs. of 2 variables
df_onehot	5 obs. of 5 variables
diabetes	758 obs. of 9 variables
one_hot	num [1:5, 1:3] 0 1 0 0 1 0 0 1 0 0 ...
passengers	891 obs. of 13 variables

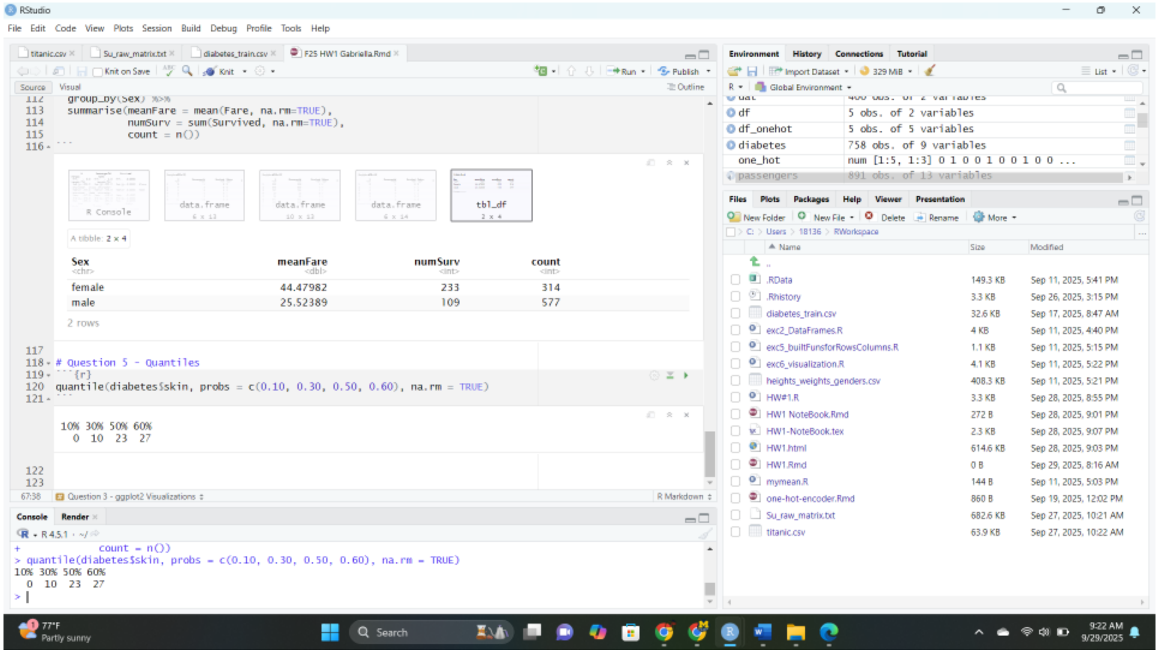
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New File
Delete
Rename
More
C: > Users > 18136 > RWorkspace

Name	Size	Modified
..		
.RData	149.3 KB	Sep 11, 2025, 5:41 PM
.Rhistory	3.3 KB	Sep 26, 2025, 3:15 PM
diabetes_train.csv	32.6 KB	Sep 17, 2025, 8:47 AM
exc2_DataFrames.R	4 KB	Sep 11, 2025, 4:40 PM
exc5_builtFunsforRowsColumns.R	1.1 KB	Sep 11, 2025, 5:15 PM
exc6_visualization.R	4.1 KB	Sep 11, 2025, 5:22 PM
heights_weights_genders.csv	408.3 KB	Sep 11, 2025, 5:21 PM
HW#1.R	3.3 KB	Sep 28, 2025, 8:55 PM
HW1 NoteBook.Rmd	272 B	Sep 28, 2025, 9:01 PM
HW1-NoteBook.tex	2.3 KB	Sep 28, 2025, 9:07 PM
HW1.html	614.6 KB	Sep 28, 2025, 9:03 PM
HW1.Rmd	0 B	Sep 29, 2025, 8:16 AM
mymean.R	144 B	Sep 11, 2025, 5:03 PM
one-hot-encoder.Rmd	860 B	Sep 19, 2025, 12:02 PM
Su_raw_matrix.txt	682.6 KB	Sep 27, 2025, 10:21 AM
titanic.csv	63.9 KB	Sep 27, 2025, 10:22 AM

77°F
Partly sunny
Search
9:22 AM
9/29/2025

Question 5 – Quantiles

- `quantile(x, probs, na.rm=TRUE)` → calculates percentiles.
- `probs = c(0.10, 0.30, 0.50, 0.60)` → 10th, 30th, 50th (median), 60th percentiles.
- `na.rm=TRUE` → ignore missing values.



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
titanic.csv
Su_raw_matrix.txt
diabetes_train.csv
F25 HW1 Gabriella.Rmd
Knit on Save
Knit
Run
Publish
Source
Visual
Outline

```r
  group_by(Sex) %>%
  summarise(meanFare = mean(Fare, na.rm=TRUE),
            numSurv = sum(Survived, na.rm=TRUE),
            count = n())
```
~~~

A tibble: 2 x 4

Sex <chr>	meanFare <dbl>	numSurv <int>	count <int>
female	44.47982	233	314
male	25.52389	109	577

2 rows

~~~r

# Question 5 - Quantiles
```{r}
quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
~~~
```

```
10% 30% 50% 60% 
  0  10  23  27 
```

67:38
Question 3 - ggplot2 Visualizations
R Markdown
Console
Render
R 4.5.1 · ~/

```
+             count = n())
> quantile(diabetes$skin, probs = c(0.10, 0.30, 0.50, 0.60), na.rm = TRUE)
10% 30% 50% 60% 
  0  10  23  27 
> 
```

Environment
History
Connections
Tutorial
Import Dataset
329 MiB
List
R
Global Environment

dat	400 obs. of 2 variables
df	5 obs. of 2 variables
df_onehot	5 obs. of 5 variables
diabetes	758 obs. of 9 variables
one_hot	num [1:5, 1:3] 0 1 0 0 1 0 0 1 0 0 ...
passengers	891 obs. of 13 variables

Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New File
Delete
Rename
More
C: > Users > 18136 > RWorkspace

Name	Size	Modified
..		
.RData	149.3 KB	Sep 11, 2025, 5:41 PM
.Rhistory	3.3 KB	Sep 26, 2025, 3:15 PM
diabetes_train.csv	32.6 KB	Sep 17, 2025, 8:47 AM
exc2_DataFrames.R	4 KB	Sep 11, 2025, 4:40 PM
exc5_builtFunsforRowsColumns.R	1.1 KB	Sep 11, 2025, 5:15 PM
exc6_visualization.R	4.1 KB	Sep 11, 2025, 5:22 PM
heights_weights_genders.csv	408.3 KB	Sep 11, 2025, 5:21 PM
HW#1.R	3.3 KB	Sep 28, 2025, 8:55 PM
HW1 NoteBook.Rmd	272 B	Sep 28, 2025, 9:01 PM
HW1-NoteBook.tex	2.3 KB	Sep 28, 2025, 9:07 PM
HW1.html	614.6 KB	Sep 28, 2025, 9:03 PM
HW1.Rmd	0 B	Sep 29, 2025, 8:16 AM
mymean.R	144 B	Sep 11, 2025, 5:03 PM
one-hot-encoder.Rmd	860 B	Sep 19, 2025, 12:02 PM
Su_raw_matrix.txt	682.6 KB	Sep 27, 2025, 10:21 AM
titanic.csv	63.9 KB	Sep 27, 2025, 10:22 AM

77°F
Partly sunny
Search
9:22 AM
9/29/2025
```