BSRP Problem Set 3

your-name-here

Introduction

The goal of this problem set is to explore more in depth of the Stephens et al. paper on KRAS mutation and expression. You will bring in tools of ggplot and statistical testing to quantify the results more carefully. Then, you will use functional genomics to validate your findings.

Here is our original question.

Biological Problem <-> Computational Analysis

Biological Question	Computational analysis
Is KRAS expression higher in KRAS mutated cell lines compared to KRAS wild-type cell lines? Does this trend hold in lung, pancreas, and colon cancer cell lines?	T-test to compare means of two groups, visualized with box plots.

Assignment Formatting

Write your code in the designated code chunks provided, and write your written answers in the text region of the document. If you want to incorporate both text and code output, consider the function `cat()`. For example, `cat("The number of rows in the iris dataframe is", nrow(iris))`.

```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc
```

```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc
```

```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc
```

```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc
```

Back to Figure 1

Problem 1

Wrangle the the data back into the target dataframe of cell lines as observations, and KRAS mutation, KRAS expression, and disease type as variables. Feel free to use your code from last problem set. What is the dimension of this dataframe? This is your target dataframe for the analysis - make sure that it makes sense to you before you run your analysis!

```
## The dimension of the target dataset is 1811 29
```

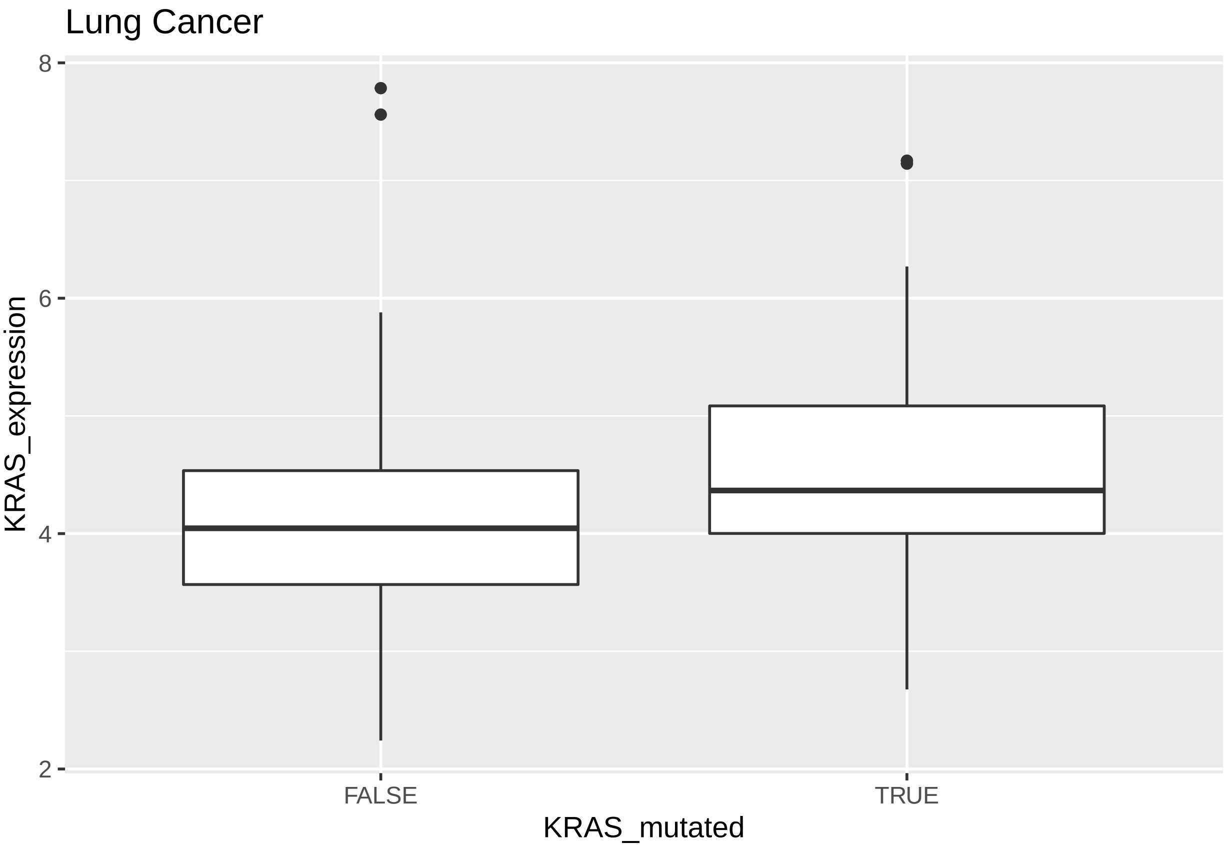
Probelm 2

Using ggplot, create the Figure 1 figure for lung cancer. Then, run a T-test and interpret the results in your own words: what is the null hypothesis, and what is the population parameter? What is your inference about the population parameter and confidence interval? How do you interpret the p-value? Also, compute the sample size used for each MT and WT groups - we will be looking at this more closely soon.

Note: When running the T-test, it has the signs flipped - a negative effect size indicates higher expression in the mutant group.

Repeat this for pancreatic cancer and colorectal cancers.

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```



```
##
##  Welch Two Sample t-test
##
```

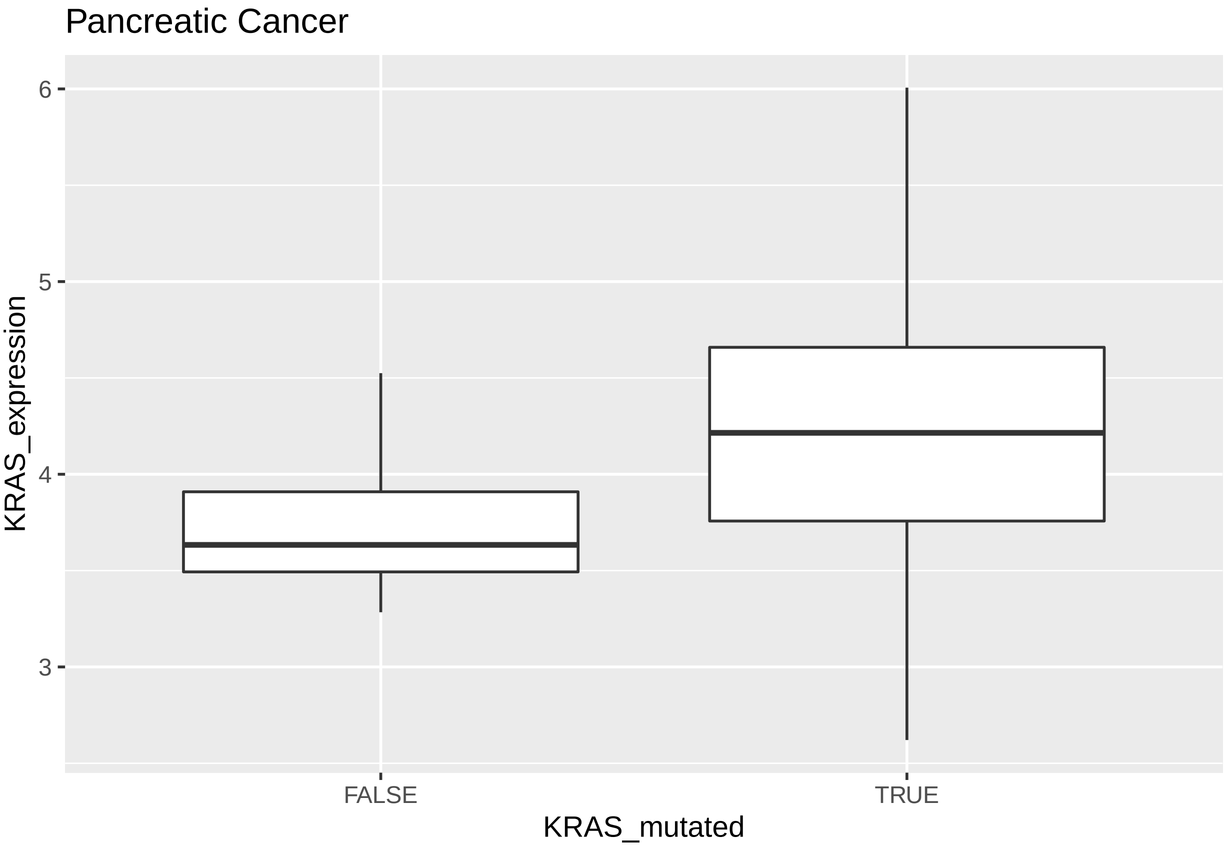
```
## data:  KRAS_expression by KRAS_mutated
## t = -3.1773, df = 68.276, p-value = 0.002233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7846729 -0.1793052
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.082042            4.564031
```

```
## Sample size of MT 51
```

```
## Sample size of WT 222
```

Solution for lung cancer: The null hypothesis is that there is no difference in means between the two groups. We inferred that the difference between the two groups is 4.06 - 4.53 = -.47 favoring higher expression in mutant with a 95% confidence interval of (-.77, -.16). The p-value is < .05, so we reject the null hypothesis, favoring that there is a difference in mean gene expression.

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```



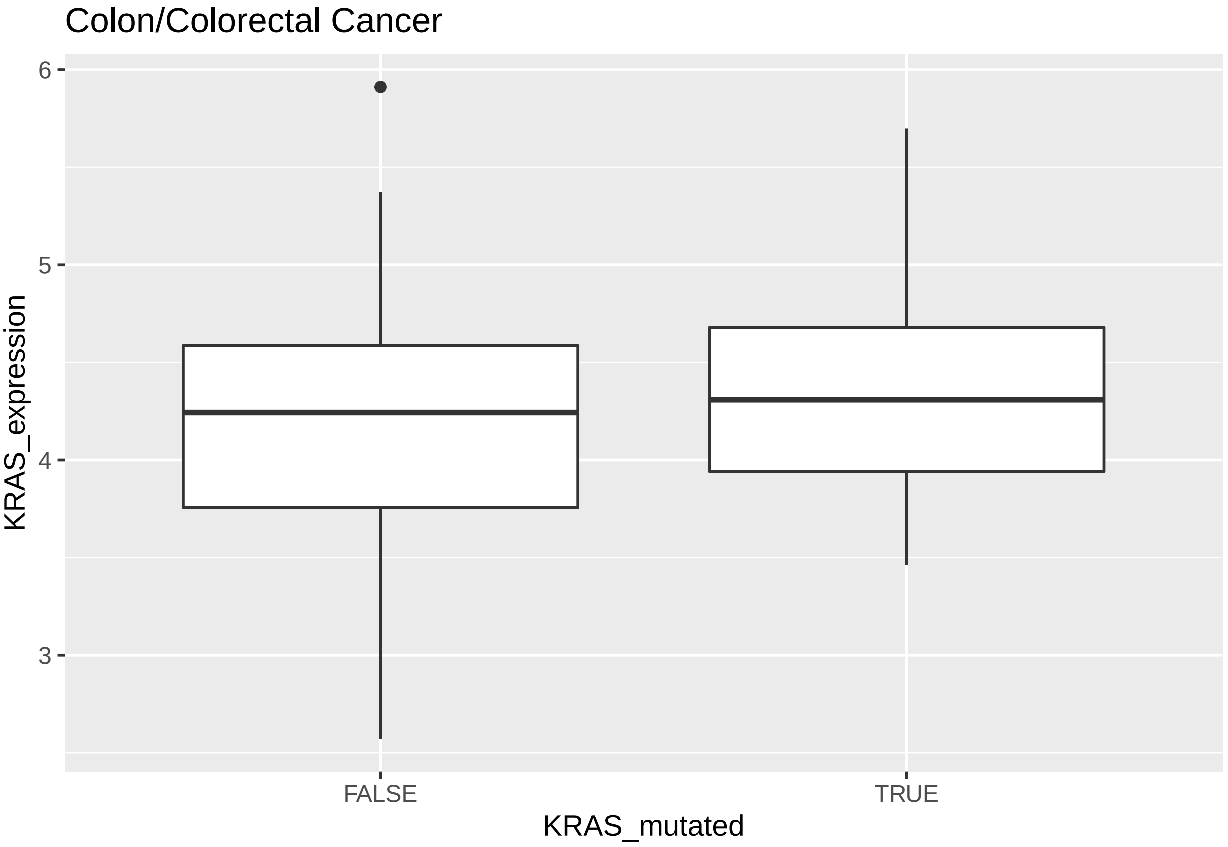
```
##
##  Welch Two Sample t-test
##
## data:  KRAS_expression by KRAS_mutated
## t = -1.7536, df = 3.9862, p-value = 0.1546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.298446  0.294007
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3.768915            4.271135
```

```
## Pancreatic cell lines: Sample size of MT 52
```

```
## Pancreatic cell lines: Sample size of WT 7
```

Solution for pancreatic cancer: The null hypothesis is that there is no difference in means between the two groups. We inferred that the difference between the two groups is 4.06 - 4.21 = -.15 favoring higher expression in mutant with a 95% confidence interval of (-1.13, .82). The confidence interval overlaps the null hypothesis of difference of 0. The p-value is > .05, so cannot reject the null hypothesis, suggesting that the difference in gene expression is minimal to none.

```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```



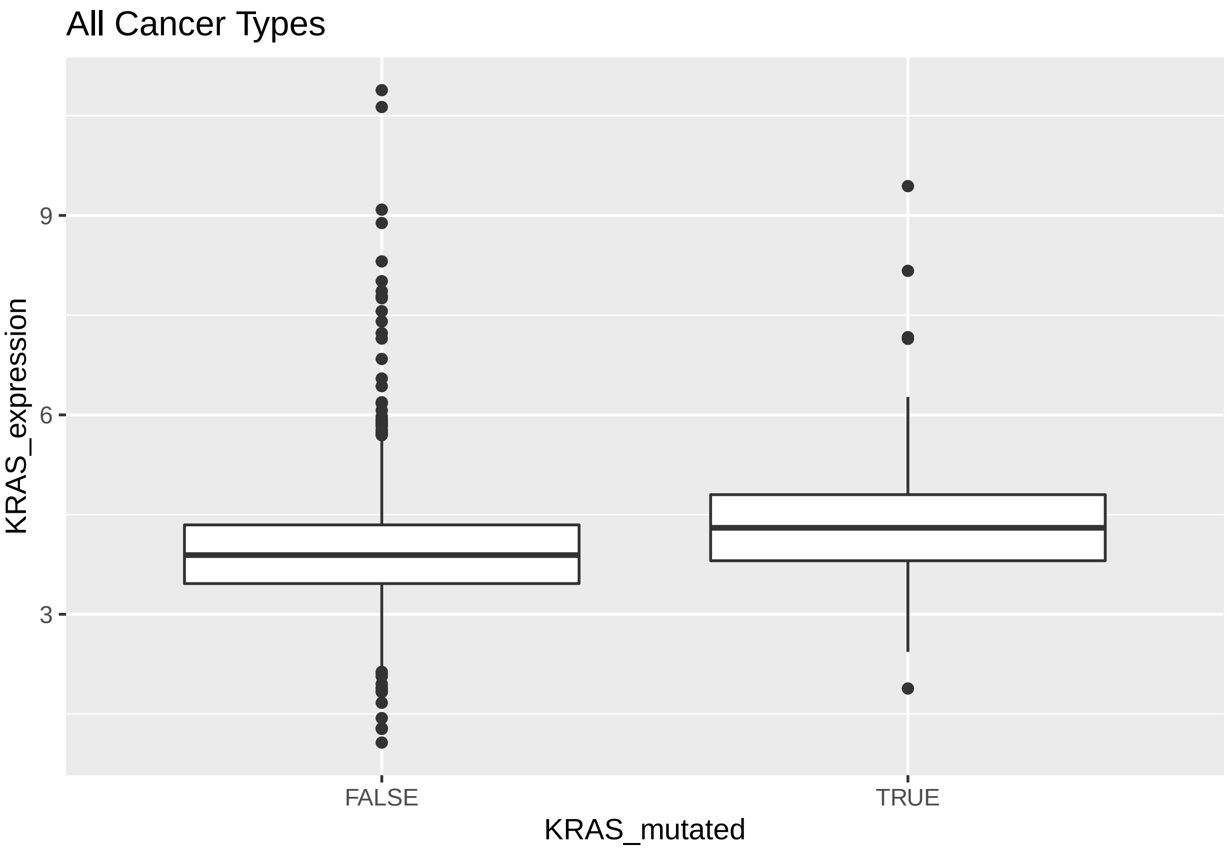
```
##
##  Welch Two Sample t-test
##
## data:  KRAS_expression by KRAS_mutated
## t = -0.83155, df = 50.299, p-value = 0.4096
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4432275  0.1836579
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.176838            4.306623

## Colon cell lines: Sample size of MT 44

## Colon cell lines: Sample size of WT 39
```

Solution for colon cancer: The null hypothesis is that there is no difference in means between the two groups. We inferred that the difference between the two groups is 4.17 - 4.28 = -.11 favoring higher expression in mutant with a 95% confidence interval of (-.44, .22). The confidence interval overlaps the null hypothesis of difference of 0. The p-value is > .05, so cannot reject the null hypothesis, suggesting that the difference in gene expression is minimal to none.

Problem 3

Run the same analysis of KRAS expression compared with mutation looking at *all* cell lines. Compare this analysis to one of your analysis in Problem 2: the anlaysis you performed in Problem 2 is on a subset of data of your analysis here. Do you always expect the trends of your analysis be same between the entire group and the subgroup? Hint: For a similar analysis, consider the section "KRAS expression and CNV" of the paper.

```
## Warning: Removed 436 rows containing non-finite values (stat_boxplot).
```



```
##
##  Welch Two Sample t-test
##
## data:  KRAS_expression by KRAS_mutated
## t = -6.0145, df = 312.12, p-value = 5.043e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5202808 -0.2637812
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3.955292            4.347323
```

```
## Sample size of MT 245
```

```
## Sample size of WT 1566
```

It still trends in the positive direction that mutant KRAS is associated with higher KRAS expression. However, subgroups of this does not necessarily have to be true - some subtypes of cancer can have lower KRAS expression when KRAS is mutated, despite the global trend shown here.

Functional Genomics

Think back to our original hypothesis: mutant, oncogenetic KRAS leads to uncontrollable cell growth. We have shown that mutant KRAS cell lines exhibit more KRAS expression, hence more KRAS protein activity, associated with cancer growth. This is an association. How can we be more sure that the cancer cell *needs* mutant KRAS to proliferate? What happens if we take a cancer celline with KRAS mutation, and inactivate the gene? If most of the cell line dies from such an intervention, then we know it may have a large impact on the survival of the cell line. The mutation is *necessary* for the cancer's survival. Furthermore, if corresponding cell lines without KRAS mutation survives from such an intervention, then we have identified a genotype essential for cancer cells but not non-KRAS mutation cells.

The *Dependency Map* CCLE cell lines, deleted a gene using CRISPR, and measured how well the cell line survived that intervention. This process was repeated for almost every human gene. The resulting dataset, the Dependency Map, can be explored!

A few terms to be defined:

A cell line is *dependent* on gene X if *knocking out* gene X by CRISPR causes the cell line to reduce its ability to grow.

The ability of a cell line to grow is called its *viability*. When a cell line's viability decreases, it is going through *depletion*. When a cell line's viability increases, it is going through *proliferation*.

One needs some controls when analyzing a CRISPR experiment. An *essential gene* is a gene that is necessarily for all cells to survive. A *nonessential gene* is a gene that is not necessarily for all cells to survive.

Biological Problem <-> Computational Analysis

Biological Question	Computational analysis
Is KRAS mutation associated with strong KRAS dependency?	T-test to compare means of two groups, visualized with box plots.

Exploring Dependency Map Data

The dependency score is calculated in the following:

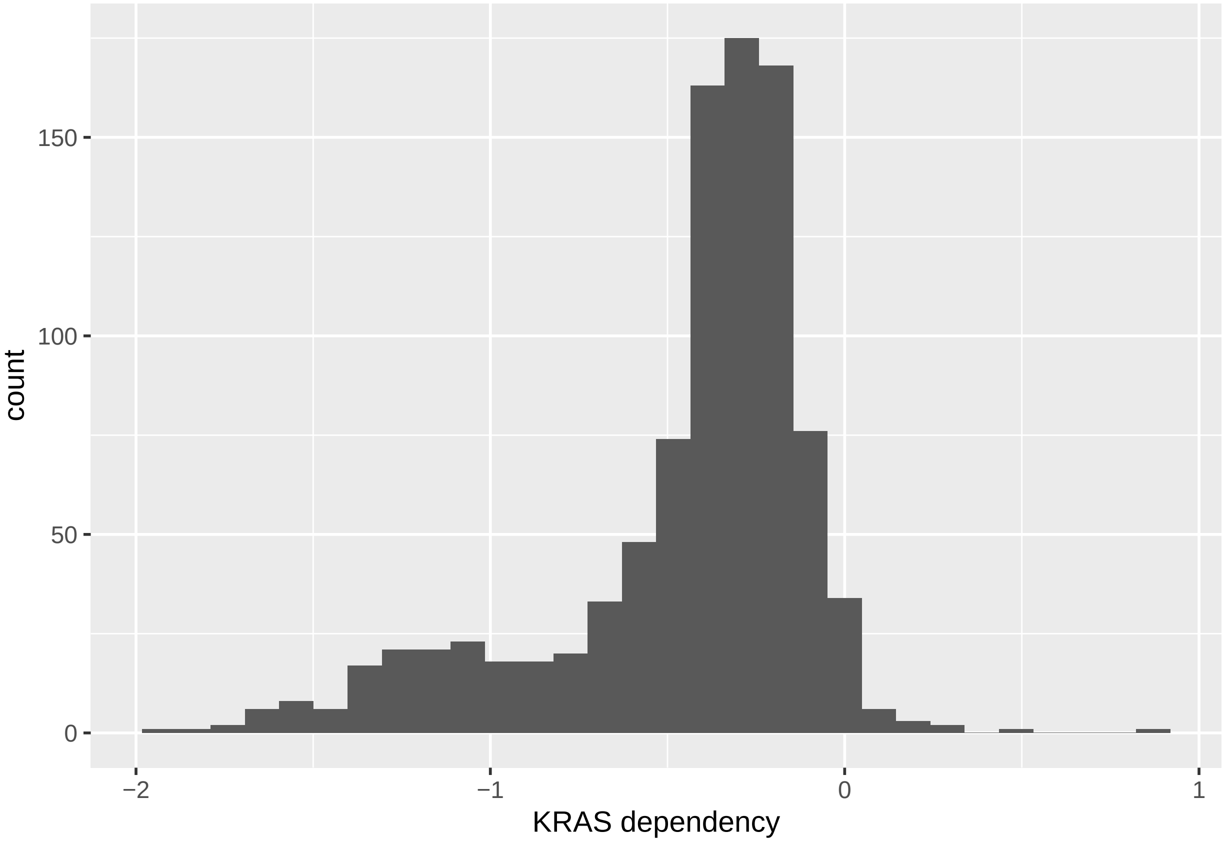
- 0 indicates no effect: the number of cells after the CRISPR knock-out is similar to the effect of knocking out nonessential genes. The median dependency score of a list of nonessential genes is 0.
- Negative score indicates depletion: the number of cells after CRISPR knock-out has decreased. The median dependency score of a list of essential genes is -1. Most cells will have died if the score is -1.
- Positive score indicates proliferation: the number of cells after CRISPR knock-out has increased.

Problem 4

Run a few commands to get to know the `dependency` data. It is structured the most similar to the expression data. How many genes and cell lines are profiled? Then, using ggplot, create a histogram of KRAS dependency, and give some summary statistics about the dependency score. What can you conclude about KRAS dependency in this collection of cell lines?

```
## Dependency has 946 cell lines.
```

```
## Dependency has 17646 genes.
```



```
## The mean dependency score is: -0.444865
```

```
## The median dependency score is: -0.3436008
```

Solution: For most cell lines, knocking out KRAS has a killing effect on the cell lines.

Problem 5

Wrangle a new target dataframe with the following observations and variables:

Observations: cell lines.

Variables: KRAS mutated, KRAS expression, KRAS dependency, and the Primary Disease.

Hint: working on top of what you have in Problems 1-2 will make things easier. Warning: The columns for KRAS in expression and dependency are both named `KRAS`, so this is going to be confusing if you don't rename at least one of your columns to something else. Use the `rename()` function to fix this issue. For example, `dataframe = rename(dataframe, KRAS_expression = KRAS)`

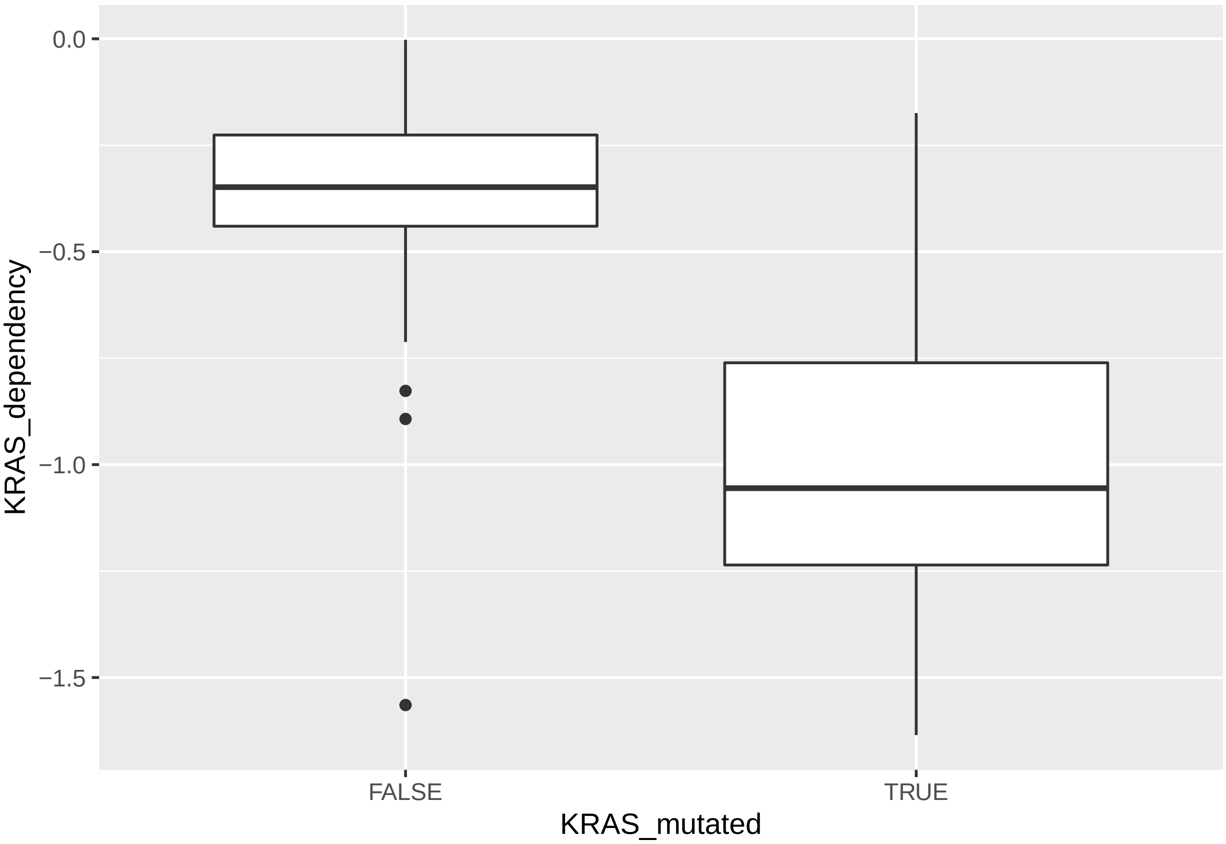
Run some checks to make sure that you got the right target dataframe! What is the dimension of this target dataframe?

```
## Dimention of dataframe: 1811 30
```

Problem 6

Perform a 2-sample T-test of dependency score between KRAS mutated and wildtype for lung cancer cell lines. Analyze it similar how you did in Problem 2. Interpret what the mean dependency score means roughly for each wildtype and mutant group.

```
## Warning: Removed 149 rows containing non-finite values (stat_boxplot).
```



```
##
##  Welch Two Sample t-test
##
## data:  KRAS_dependency by KRAS_mutated
## t = 10.639, df = 46.842, p-value = 4.353e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5312170 0.7789765
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          -0.3550092          -1.0101059
```

```
## Lung cell lines: Sample size of MT 51
```

```
## Lung cell lines: Sample size of WT 222
```

Solution for lung cancer: The null hypothesis is that there is no difference in means of dependency score between the two groups. We inferred that the difference between the two groups is -0.48 - -1.167 = .69 favoring lower dependency in mutant with a 95% confidence interval of (.56, .82). The p-value is < .05, so we reject the null hypothesis, favoring that there is a difference in mean dependency score.

KRAS mutated cell lines has a mean dependency score of ~-1, indicating that KRAS-mut is similar to essential genes, while KRAS-wt is somewhere between essential and nonessential genes.

Problem 7

Based on what you found in Problem 5, how would you conclude your findings? What are the implications, and what analysis or experiments would you follow up with?

Solution: KRAS mutation is an essential genotype for cancer dependency in lung cancer cell lines, when compared to KRAS wildtype.

In the context of precision medicine, this leads to the hypothesis that if a patient has lung cancer and KRAS mutation, then their cancer cells are highly dependent on KRAS for growth. Having a drug that targets KRAS protein would lead to cancer specific cell death.

There are many things that we don't know about this cancer dependency: We are comparing against other KRAS-WT lung cancer cell lines, and a significant amount of their cells will die off also. We need to do a lot more comparisons to characterize the toxicity of such an intervention as knocking out KRAS: comparison to healthy cells, a range of model systems, etc. We want to be careful whether this is a cancer-specific dependency vs. tissue-specific dependency. We might also study the downstream pathway carefully to understand what knocking out KRAS is doing exactly: MAPK, PIK3CA, etc. (There is work being done now at the Dependency map to profile omics after knocking out the cell to investigate these hypothesis.)

Turns out, there have been large efforts to find ways to target KRAS but no successful drugs have been found, but that may be changing as of last year. There has been one promising drug targeting one specific, less common mutation of KRAS. What else can be done? A lot of work has been done to search for other dependencies in the context when KRAS is mutated. PIK3CA is dependent in certain context of KRAS mutants.