



COMP320

RESEARCH PRACTICE

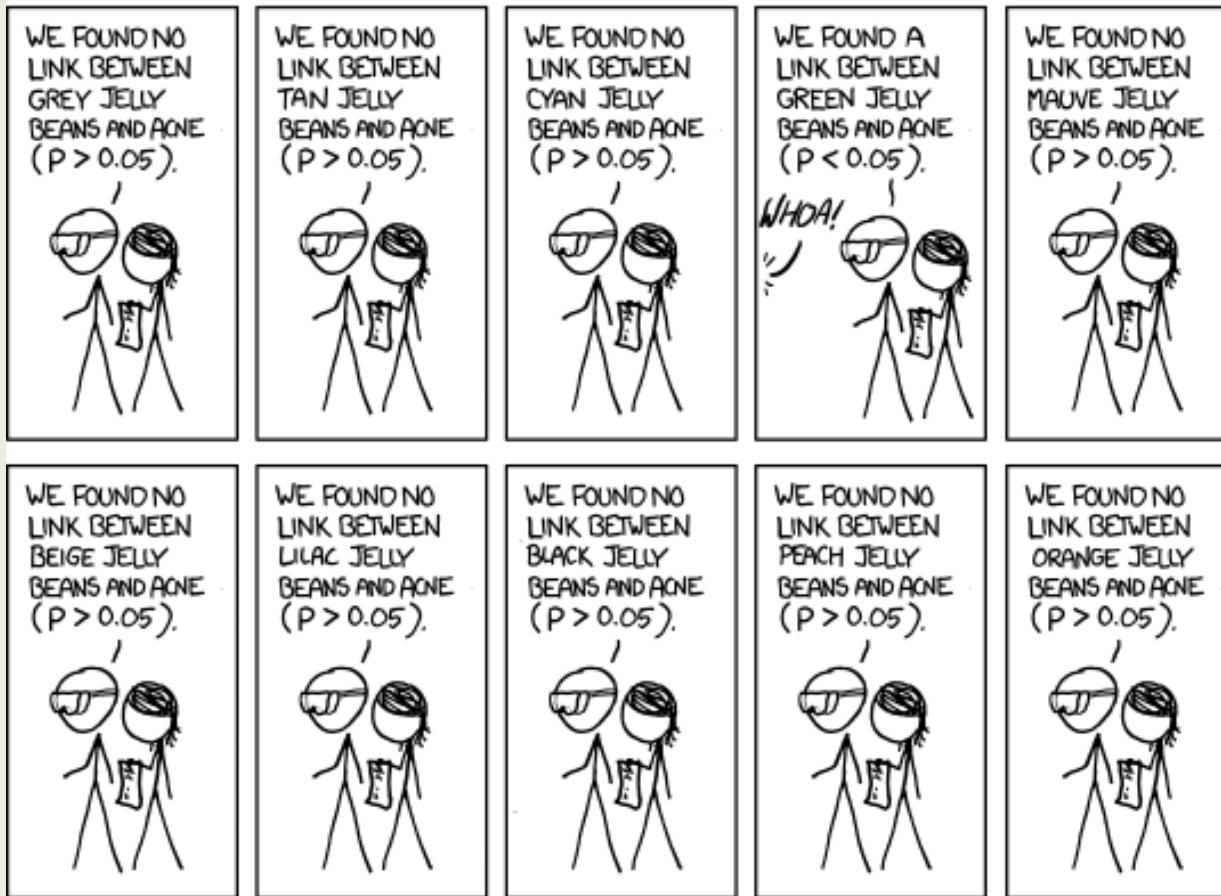
Statistical Data Analysis I

False Discovery & Replicability

When we make claims as researchers, we are often subject to error and so replication is important:

		Reality	
		True	False
Decision	Accept Claim	Good Decision	Type-1 Error ("False Positive")
	Reject Claim	Type-2 Error ("False Negative")	Good Decision

False Discovery & Replicability



False Discovery & Replicability

An alternative to replication is to use a larger dataset and to assess multiple hypotheses with reasonable adjustment. For example:

- **Benjamini Hochberg**
Used in multiple hypothesis testing on the same dataset
- **Bonferroni**
Used for multiple group comparisons using the same dataset

Common Statistical Tests

Use in your Dissertations

Research Questions

- Researchers often want to know if there is a significant **relationship** between two variables
 - *How strong is the relationship between **login queue waiting time** and **player satisfaction**?*
 - *Does an increase in **in-game prompts** correspond to an increased number of **in-game micro transactions**?*
 - *Is there a relationship between employees' **levels of stress** and **talent retention**?*
 - *Does **age** predict **play-style**?*

Research Questions

- Or if there is a **difference** between one or more scores/conditions/groups
 - *Is game-0 selling more copies than game-1 in the UK?*
 - *Does working in pairs improve programming performance?*
 - *Has allowing programmers to work from home decrease game development productivity?*
 - *Has the new character caused a significant disruption to game balance in terms of win/loss ratio?*
 - *Is there a difference in the way male and female players interact with costume-orientated micro transactions in an MMORPG?*

Statistical tests: Why we need them?

- Does **marital status** affect **play**?
- *'In a scale from 1 to 8, how often do you play games?'*
- Great! I found a difference!!

marital status	Mean	N	Std. Deviation
single	5.58	26	2.318
married/defacto	5.19	86	2.384
divorced	5.38	8	2.200
widowed	7.67	3	3.215
Total	5.34	123	2.381

- How confident are you? Is it an accident (due to chance)?
- We need to have a statistical test to make the inference!

Choosing the right test

- If you browse any introductory statistics text book you'll find a bewildering array of different statistical tests
- Each has:
 - a specific purpose (i.e. exploring relationships, comparing groups)
 - Assumptions and data requirements (categorical, ordinal or continuous data, normal distribution)
- Most of the well known tests are very easy to run in R
- However, it is critically important to be able to
 - Select the most appropriate test given your research question
 - Understand conceptually what the test is computing
 - Effectively interpret the output

Step 1: What is your question?

- Remember, when conducting research it is important to be clear about the questions you are trying to answer..ideally before you begin data collection
- The questions...
 - *Does an increase in **in-game prompts** correspond to an increased number of **in-game micro transactions**?*
 - *Is there a relationship between employees' **levels of stress** and **talent retention**?*
- ...require quite different statistical tests to questions like:
 - *Is **game-0** selling more copies than **game-1** in the UK?*
 - *Has allowing programmers to **work from home** decrease game development **productivity**?*

Step 2: Select your data

- Which variables will you be using?
- Which is the **independent variable (IV)**?
 - The variable that is believed to affect the dependent variable
 - What you control/manipulate
- Which is the **dependent variable (DV)**?
 - The observation that is believed to be affected by the IV
 - What you measure (aka outcome variable)
- Identify the IV and DV in the following questions:
 - Does gender affect product ratings?
 - Does revision time affect test scores?
 - Does the website background colour influence reading speed?
 - Which type of interface results in higher user satisfaction?

Step 2: Select your data

- What is the level of measurement for each variable?
 - Categorical or continuous?
 - Examples of categorical variables?
 - Examples of continuous variables?

Step 3: Describe your data

- Descriptive statistics should be used to define the characteristics of your data (last lecture)
- For categorical variables you need to know if numbers in each group/category are balanced
 - (e.g. Reliable comparison of gender effect not possible if 25 males and only 3 females)
- For continuous variables you need to know if the distribution is normally distributed (e.g. Not skewed)

Relationships vs. Difference

- Analysing relationships:
 - **Correlation** – are continuous variables, X and Y, related?
 - Pearson's rho for normally-distributed ratio data
 - Spearman's rank correlation for non-normal or interval data
 - **Chi-square** – is there an association between two categorical variables.
 - Useful for inferring differences between groups on discrete measures
 - Also used for 'goodness of fit' tests when comparing matrices
 - **Regression** – does the 'level' of X predict the 'level' of Y?
 - OLS regression for continuous data with normally distributed residuals
 - Logistic regression used for discrete data, based on probability of belonging
 - Flexible, can re-code some nominal/ordinal data as binary values

Relationships vs. Difference

Analysing differences:

- **T-tests** - differences between two groups (e.g. experienced, inexperienced) according to some *continuous variable* (e.g. score)
- **Mann-Whitney U Test** – based on ranks, lower power but more robust and can compare two groups with non-normal data.
- **Analysis of Variance** (ANOVA) measure differences when there are *more than two groups*.
- **Kruskal-Wallis H Test** – like Mann-Whitney U, but for multiple groups.
- **Analysis of Co-variance** (ANCOVA) measure differences when there are *more than two groups and/or continuous predictors*.
 - Essentially, combines ANOVA with regression.
- **MANOVA/MANCOVA – multivariate versions** for more than one dependent variable.

Import Data Into R

Prepare for Analysis

Import Data into R

Download and examine

https://www.dropbox.com/s/6x44olpr3kwdkkh/obfuscated_data.csv?dl=1

Import Data into R

To import data from a pre-prepared CSV file, use the following command. Note: Requires the Rcpp module. Run Rstudio in admin mode to install if not available.

```
> library(readr)  
> dat <- read_csv("E:/Stats/obfuscated_data.csv")
```

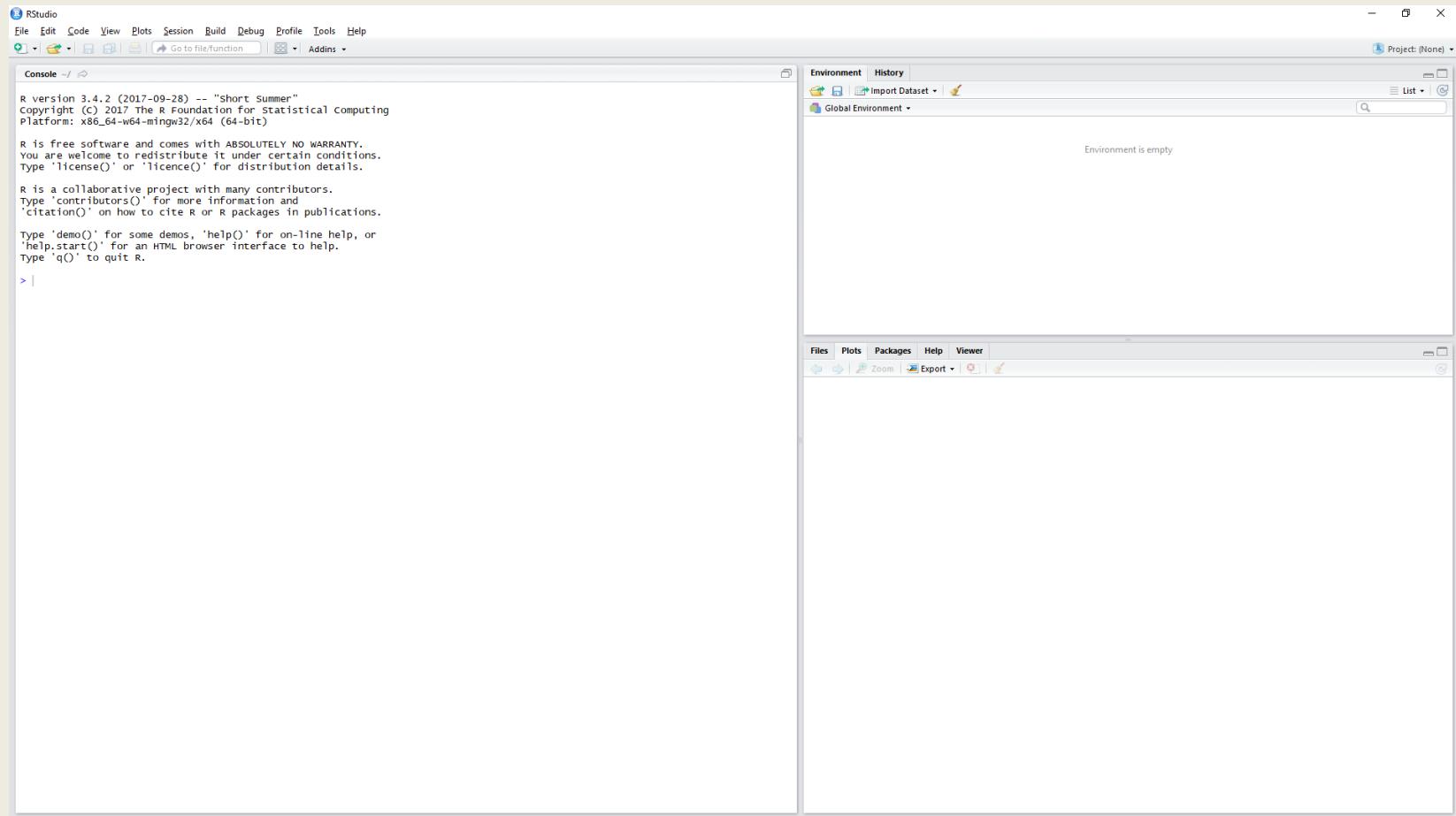
To view the data use:

```
> view(dat)
```

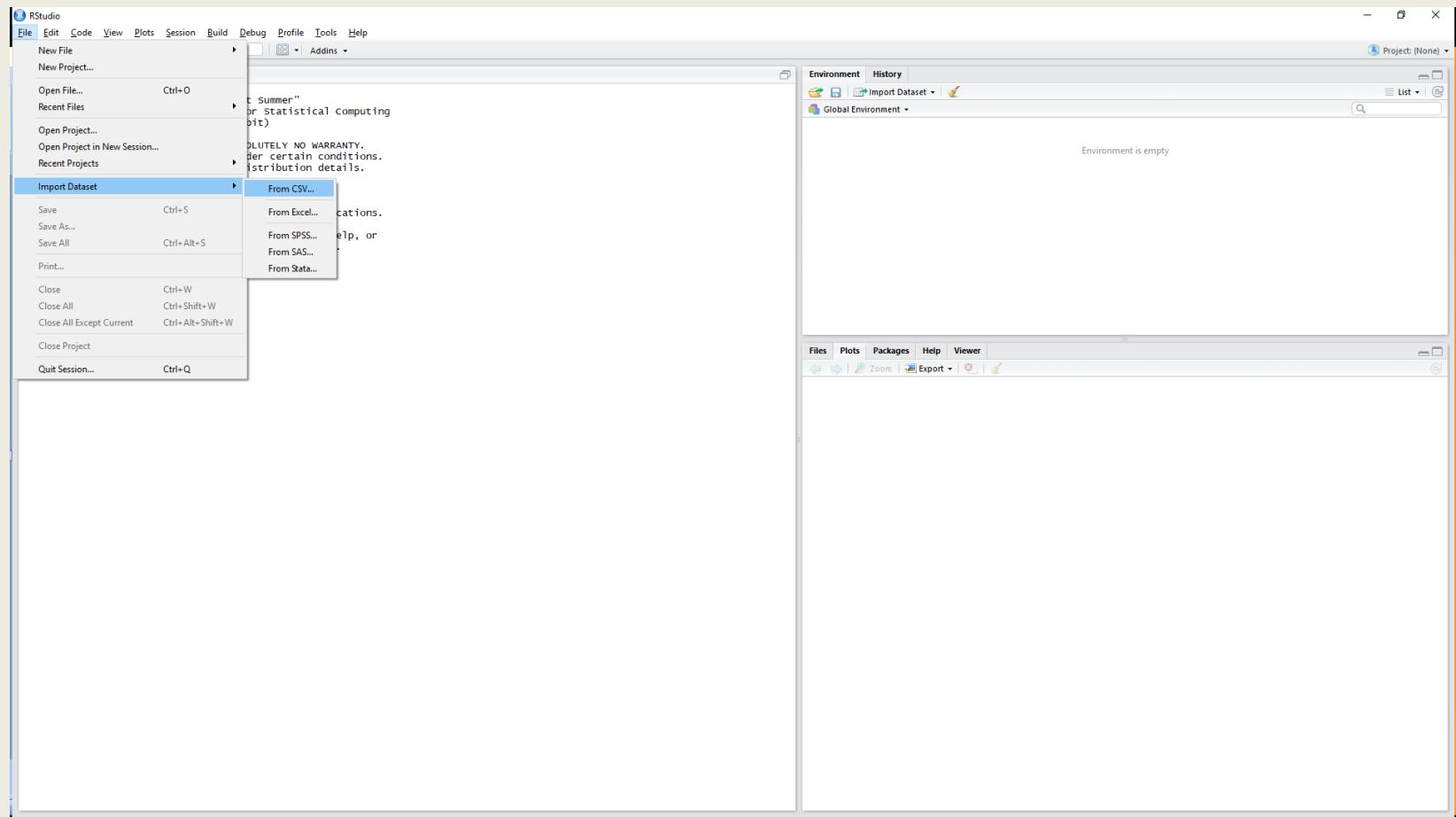
To see the descriptive statistics for all of the variables in the data use:

```
> summary(dat)
```

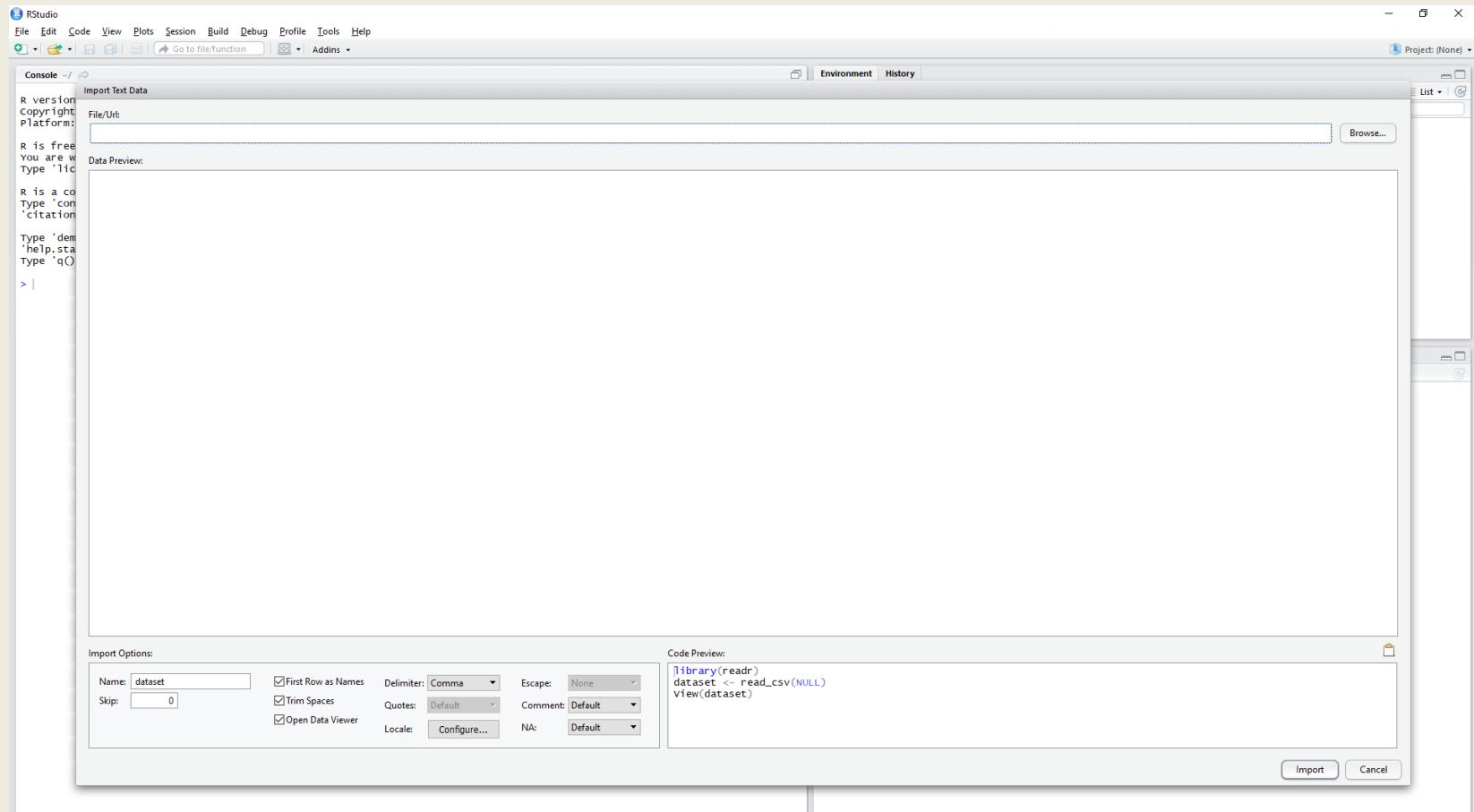
Import Data into R



Import Data into R



Import Data into R



Import Data into R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

File/Unit E:/Stats/obfuscated_data.csv

Project: (None)

Console Import Text Data

R version 4.2.1 (2022-09-22 ucrt) Copyright (C) 2022 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32 mingw-vita

R is free You are welcome to redistribute it under certain conditions.

R is a complete system for statistical computation and graphics.

Type 'citation()' to see the citation info.

Type 'demo()' to see some demos.

Type 'help()' to get help on a specific topic.

Type 'q()' to quit R.

Data Preview:

GENDER	BIRTH_YEAR	PRIOR_EXP	DEPTH	PROCAST	APTITUDE	CONCEPT	ANXIETY	INTEREST	EXTRAVERSION	AGREEABLENESS	CONSCIENTIOUSNESS	NEUROTICISM	OPENNESS
1	1988	1	64.32026	44.3969512	52.4157313	-9.0159287	55.8639910	45.80070	159	114	123	94	115
1	1997	2	57.85945	6.7252081	17.0357834	16.3427047	12.7963314	87.81904	90	125	125	90	190
1	1997	3	69.83570	-5.9635691	1.1630303	21.7374076	0.8885674	74.34215	129	105	117	67	196
1	1997	1	95.01274	8.42823205	3.0450553	27.5326928	1.1371675	91.05812	125	200	200	0	100
1	1998	2	42.83400	8.5835718	2.9886561	14.7816119	12.7550108	87.72878	135	190	155	75	170
1	1997	3	58.28550	-5.3574882	1.4622916	22.9424523	2.7831167	76.15295	159	97	109	43	200
1	1997	1	60.22861	22.7172682	16.3874860	-0.9975156	58.0333459	74.65679	58	115	97	153	177
1	1998	3	41.96357	9.885579	12.9877121	6.4188561	42.3085330	77.87651	110	90	165	150	90
1	1996	2	63.54526	15.1506880	44.2646971	20.6250292	7.4959895	86.93600	74	36	125	60	39
1	1996	1	71.11964	1.0816920	23.3052026	14.2983597	44.4814905	76.80222	188	92	160	110	101
1	1998	3	41.56587	69.6190892	74.8049680	-34.5057000	71.1038589	45.28539	153	70	80	98	106
1	1999	1	31.98191	9.0439138	5.3655622	22.0194601	6.8136457	83.81599	0	150	110	5	40
1	1997	3	68.55234	45.6862565	44.9627373	-22.4635138	76.7540448	67.50218	70	160	90	105	100
2	1995	1	75.17277	18.6474628	25.2389340	-7.0767733	26.4382562	80.20217	20	140	150	170	120
1	1997	2	35.32151	-4.0963788	-0.2446349	20.9037411	1.1786560	65.78225	150	119	167	39	113
1	1998	1	36.75300	20.8437453	7.2415136	9.4860682	22.3418757	73.26496	100	90	60	120	155
1	1991	3	54.11495	30.8031366	32.8769430	-2.3963526	47.8763479	74.80755	37	130	107	162	136
1	1993	1	62.14319	-3.7220513	1.4288541	11.8524059	6.5351831	60.93766	140	110	170	95	105
1	1998	1	40.18953	2.6112882	17.9931440	19.1064163	8.9008071	80.90829	31	159	145	50	100
1	1997	1	96.68076	-3.1001009	6.5099150	24.8916803	8.2616789	90.76659	31	22	134	199	162
1	1997	1	74.79789	8.3137345	1.6584391	16.0328773	6.5716026	88.84695	100	105	80	30	115
1	1997	3	46.70051	18.7622800	23.3187545	-1.8225908	60.4589703	70.17935	57	125	97	157	145
1	1997	3	85.13806	5.032124	12.3084754	8.0657198	16.3727413	85.27187	125	180	180	125	200
1	1997	1	46.41362	18.9195415	32.1414625	-6.2150047	50.5409066	65.37413	55	120	160	180	80

Previewing first 50 entries.

Import Options:

Name: obfuscated_data	<input checked="" type="checkbox"/> First Row as Names	Delimiter: Comma	Escape: None
Skip: 0	<input checked="" type="checkbox"/> Trim Spaces	Quotes: Default	Comment: Default
<input checked="" type="checkbox"/> Open Data Viewer	Locale: Configure...	NA: Default	

Code Preview:

```
library(readr)
obfuscated_data <- read_csv("E:/Stats/obfuscated_data.csv")
View(obfuscated_data)
```

Import Cancel

Import Data into R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

File/Unit: E:/Stats/obfuscated_data.csv

Project: (None)

Console Import Text Data

R version 4.2.1 (2022-09-22 ucrt) Copyright (C) 2022 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32 mingw-vita

R is free You are welcome to redistribute it under certain conditions.

R is a core component of the R system. Type 'citation("R")' for more information.

Type 'demo()' to see what is available in R. Type 'help.start()' to get help from your local R site. Type 'q()' to quit R.

Import Options:

Name: <input type="text" value="dat"/>	<input checked="" type="checkbox"/> First Row as Names	Delimiter: <input type="button" value="Comma"/>	Escape: <input type="button" value="None"/>
Skip: <input type="text" value="0"/>	<input checked="" type="checkbox"/> Trim Spaces	Quotes: <input type="button" value="Default"/>	Comment: <input type="button" value="Default"/>
	<input checked="" type="checkbox"/> Open Data Viewer	Locale: <input type="button" value="Configure..."/>	NA: <input type="button" value="Default"/>

Code Preview:

```
library(readr)
obfuscated_data <- read_csv("E:/Stats/obfuscated_data.csv")
View(obfuscated_data)
```

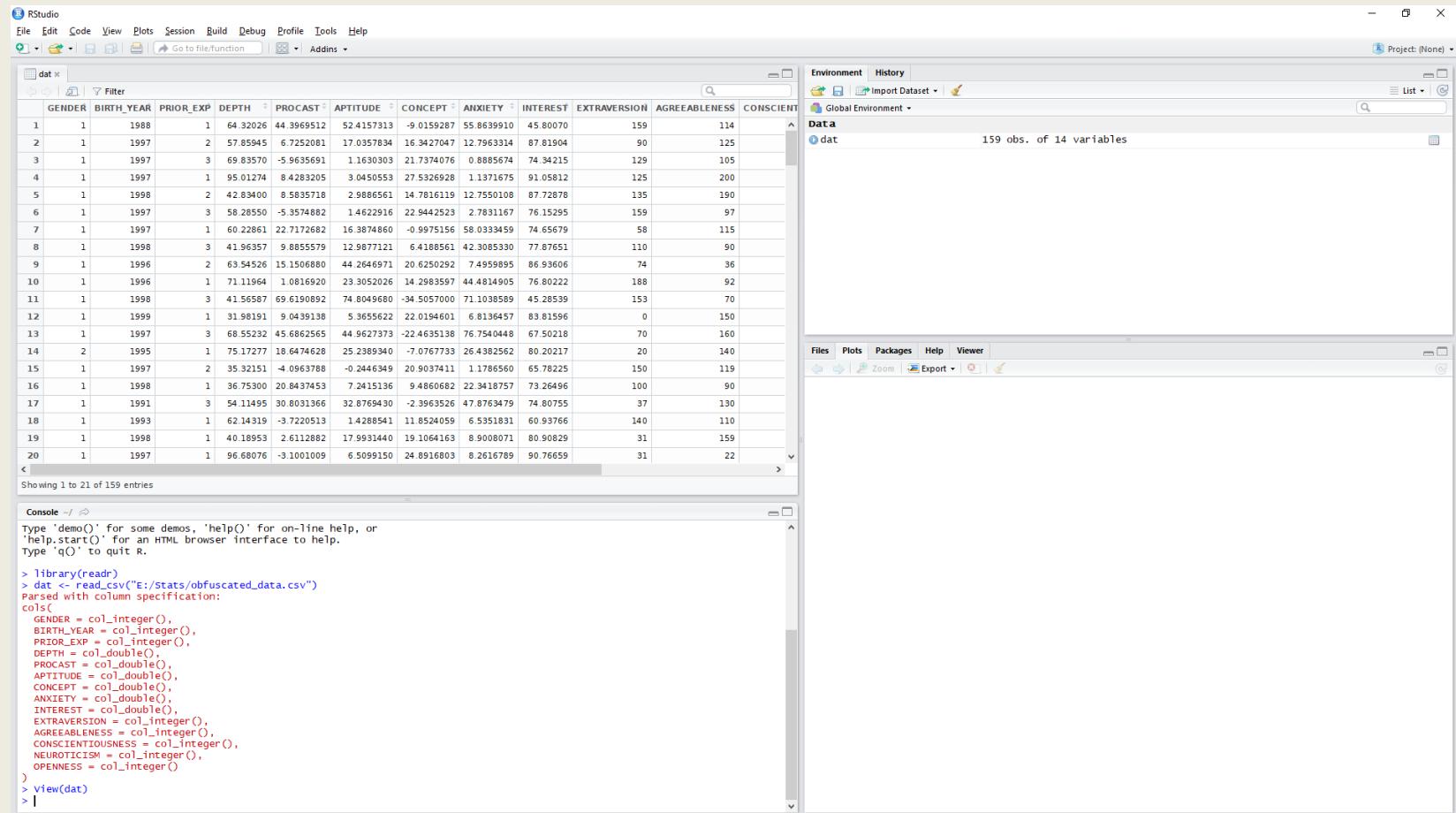
Import Cancel

Data Preview:

GENDER (integer)	BIRTH_YEAR (integer)	PRIOR_EXP (integer)	DEPTH (double)	PROCAST (double)	APTITUDE (double)	CONCEPT (double)	ANXIETY (double)	INTEREST (double)	EXTRAVERSION (integer)	AGREEABLENESS (integer)	CONSCIENTIOUSNESS (integer)	NEUROTICISM (integer)	OPENNESS (integer)
1	1988	1	64.32026	44.3969512	52.4157313	-9.0159287	55.8639910	45.80070	159	114	123	94	115
1	1997	2	57.85945	6.7252081	17.0357834	16.3427047	12.7963314	87.81904	90	125	125	90	190
1	1997	3	69.83570	-5.9635691	1.1630303	21.7374076	0.8885674	74.34215	129	105	117	67	196
1	1997	1	95.01274	8.42828205	3.0450553	27.5326928	1.1371675	91.05812	125	200	200	0	100
1	1998	2	42.83400	8.5835718	2.9886561	14.7816119	12.7550108	87.72878	135	190	155	75	170
1	1997	3	58.28550	-5.3574882	1.4622916	22.9442523	2.7831167	76.15295	159	97	109	43	200
1	1997	1	60.22861	22.7172682	16.3874860	-0.9975156	50.0333459	74.65679	58	115	97	153	177
1	1998	3	41.96357	9.8855797	12.9877121	6.4188561	42.3085330	77.87651	110	90	165	150	90
1	1996	2	63.54526	15.1506880	44.2646971	20.6250292	7.4959895	86.93606	74	36	125	60	39
1	1996	1	71.11964	1.0816920	23.3052026	14.2983597	44.4814905	76.80222	188	92	160	110	101
1	1998	3	41.56587	69.6190892	74.8049680	-34.5057000	71.1038589	45.28539	153	70	80	98	106
1	1999	1	31.98191	9.0439138	5.3655622	22.0194601	6.8136457	83.81599	0	150	110	5	40
1	1997	3	68.55234	45.6862565	44.9627373	-22.4635138	76.7540448	67.50218	70	160	90	105	100
2	1995	1	75.17277	18.6474628	25.2389340	-7.0767733	26.4382562	80.20217	20	140	150	170	120
1	1997	2	35.32151	-4.0963788	-0.2446349	20.9037411	1.1786560	65.78225	150	119	167	39	113
1	1998	1	36.75300	20.8437453	7.2415136	9.4860682	22.3418757	73.26496	100	90	60	120	155
1	1991	3	54.11495	30.8031366	32.8769430	-2.3963526	47.8763479	74.80755	37	130	107	162	136
1	1993	1	62.14319	-3.7220513	1.4288541	11.8524059	6.5351831	60.93766	140	110	170	95	105
1	1998	1	40.18953	2.6112882	17.9931440	19.1064163	8.9008071	80.90829	31	159	145	50	100
1	1997	1	96.68076	-3.1001009	6.5099150	24.8916803	8.2616789	90.76659	31	22	134	199	162
1	1997	1	74.79789	8.3137345	1.6584391	16.0328773	6.5716026	88.84695	100	105	80	30	115
1	1997	3	46.70051	18.7622800	23.3187545	-1.8225908	60.4589703	70.17935	57	125	97	157	145
1	1997	3	85.13806	5.032124	12.3084754	8.0657198	16.3727413	85.27187	125	180	180	125	200
1	1997	1	46.41362	18.9195415	32.1414625	-6.2150047	50.5409066	65.37413	55	120	160	180	80

Previewing first 50 entries.

Import Data into R



The screenshot shows the RStudio interface with the following components:

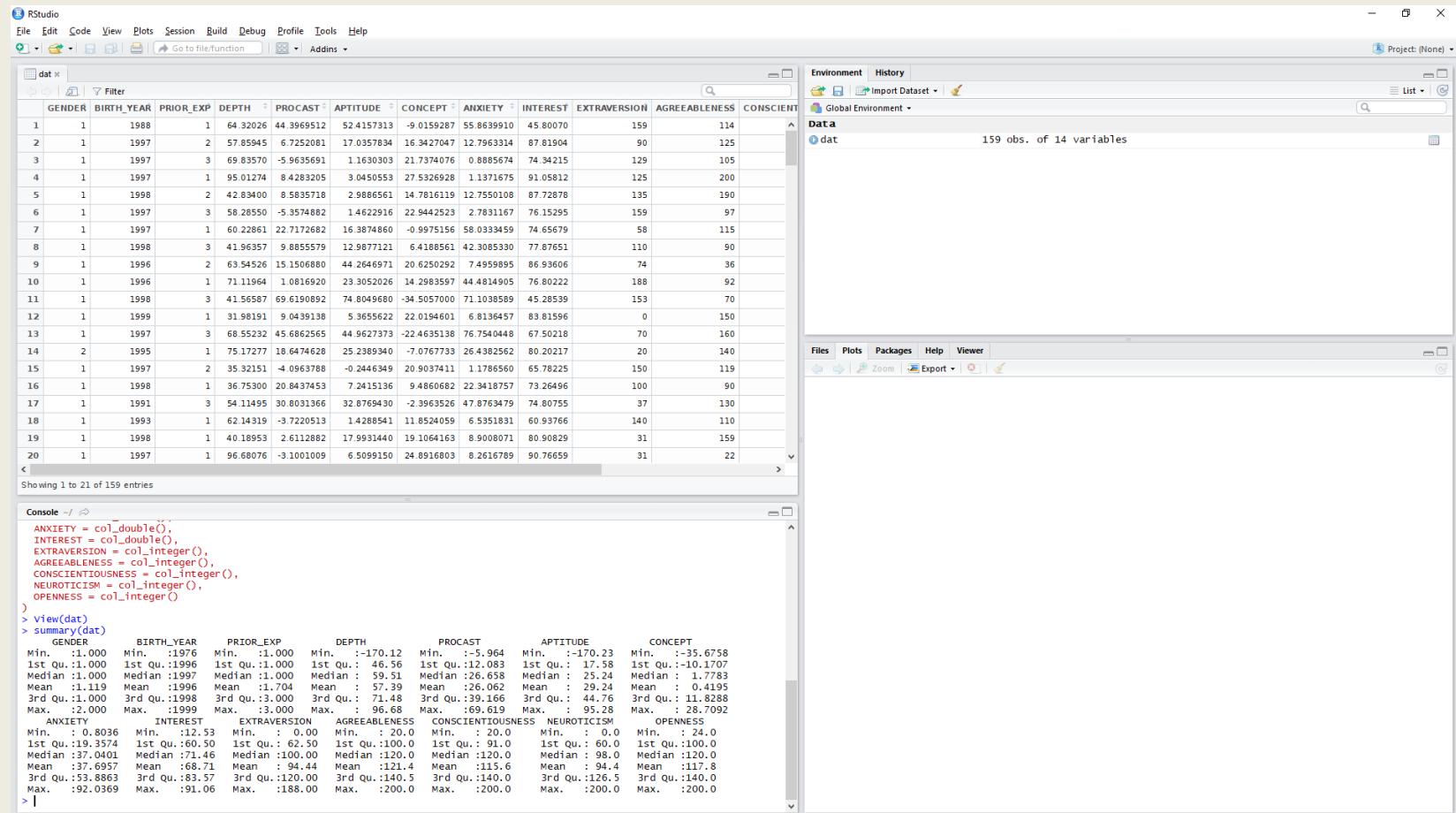
- Environment View:** Shows a data frame named "dat" with 159 observations and 14 variables.
- Data View:** Displays the first 21 rows of the "dat" data frame.
- Console View:** Shows the R code used to import the data from a CSV file.

```

dat <- read_csv("E:/Stats/obfuscated_data.csv")
Parsed with column specification:
cols(
  GENDER = col_integer(),
  BIRTH_YEAR = col_integer(),
  PRIOR_EXP = col_integer(),
  DEPTH = col_double(),
  PROCAST = col_double(),
  APTITUDE = col_double(),
  CONCEPT = col_double(),
  ANXIETY = col_double(),
  INTEREST = col_double(),
  EXTRAVERSION = col_integer(),
  AGREEABLENESS = col_integer(),
  CONSCIENTIOUSNESS = col_integer(),
  NEUROTICISM = col_integer(),
  OPENNESS = col_integer()
)
> view(dat)
>

```

Import Data into R



The screenshot shows the RStudio interface with the following components:

- File Explorer:** Shows a file named "dat" with a CSV extension.
- Environment:** Displays the dataset "dat" with 159 observations and 14 variables.
- Console:** Shows the R code used to import the data and the resulting summary statistics for each variable.

```

ANXIETY = col_double(),
INTEREST = col_double(),
EXTRAVERSION = col_integer(),
AGREEABLENESS = col_integer(),
CONSCIENTIOUSNESS = col_integer(),
NEUROTICISM = col_integer(),
OPENNESS = col_integer()
)
> view(dat)
> summary(dat)

      GENDER    BIRTH_YEAR   PRIOR_EXP    DEPTH    PROCAST    APTITUDE    CONCEPT
Min. :1.000  Min. :1976.0  Min. :1.000  Min. :-170.12  Min. :-5.964  Min. :-170.23  Min. :-35.6758
1st Qu.:1.000  1st Qu.:1996.0  1st Qu.:1.000  1st Qu.:-12.00  1st Qu.: 4.56  1st Qu.: 17.58  1st Qu.: -10.707
Median :1.000  Median :1997.0  Median :1.000  Median : 1.00  Median : 5.00  Median : 25.00  Median :  -7.933
Mean   :1.119  Mean   :1996.0  Mean   :1.000  Mean   : 7.704  Mean   : 7.39  Mean   : 25.062  Mean   : 0.41935
3rd Qu.:1.000  3rd Qu.:1998.0  3rd Qu.:1.000  3rd Qu.:-3.000  3rd Qu.: 71.48  3rd Qu.: 39.166  3rd Qu.: 11.8288
Max.  :2.000  Max.  :1999.0  Max.  :1.000  Max.  : 96.68  Max.  : 69.616  Max.  : 95.28  Max.  : 28.7092

      ANXIETY    INTEREST    EXTRAVERSION    AGREEABLENESS    CONSCIENTIOUSNESS    NEUROTICISM    OPENNESS
Min. : 0.8036  Min. :12.53  Min. : 0.00  Min. : 20.0  Min. : 0.0  Min. : 24.0
1st Qu.:19.3574 1st Qu.:60.50  1st Qu.: 62.50  1st Qu.:100.0  1st Qu.: 91.0  1st Qu.: 60.0  1st Qu.:100.0
Median :37.0401  Median :71.46  Median :100.00  Median :120.0  Median :120.0  Median :98.0  Median :120.0
Mean  :37.6957  Mean  :68.71  Mean  :94.44  Mean  :121.4  Mean  :115.6  Mean  :94.4  Mean  :117.8
3rd Qu.:53.8863 3rd Qu.:83.57  3rd Qu.:120.00  3rd Qu.:140.5  3rd Qu.:140.0  3rd Qu.:126.5  3rd Qu.:140.0
Max. :92.0369  Max. :91.06  Max. :188.00  Max. :200.0  Max. :200.0  Max. :200.0  Max. :200.0
  
```

Data Analysis in R

Illustrating Your Findings

Data Analysis in R

To access a specific variable, use the \$ character. It works a bit like the dot operator:

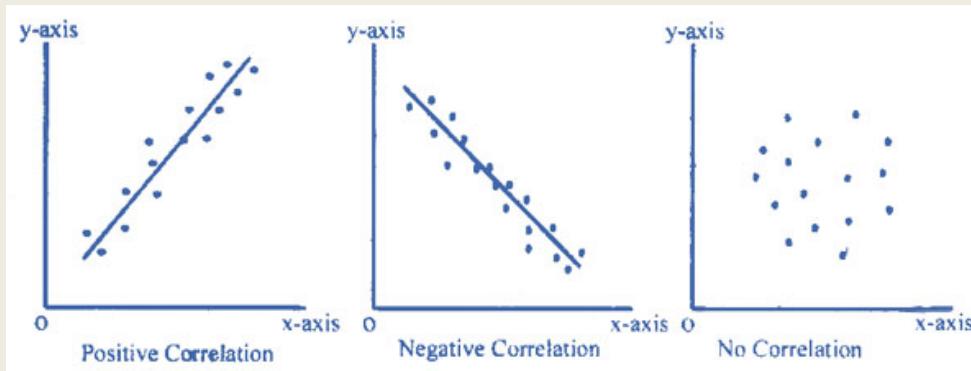
```
> describe(dat$PROCAST)
```

Note, some functions such as describe require external libraries to be installed and loaded:

```
> install.packages("psych")
> library(psych)
```

Correlation

- Extent to which two continuous variables co-vary (change together)
 - Ice-cream sales relate to temperature
 - Waiting time relates to customer satisfaction
- Correlations have
 - Direction
 - **Positive** (as the one increases the other variable increases as well)
 - **Negative** (as the one increases the other one decreases)
 - Magnitude – how closely related?



Pearson's correlation

- Provides a numerical measure of the magnitude/direction of the correlation known as r
 - any value between **-1 to +1**
 - -1 (negative), 1 (positive), Close to 0 (no/low)
- Pearson r , is a parametric test so assumes both variables
 - Are continuous
 - Have an approximately normal distribution

Correlation

To see a correlation:

```
> cor(x, y)  
> cor.test(x, y, method)
```

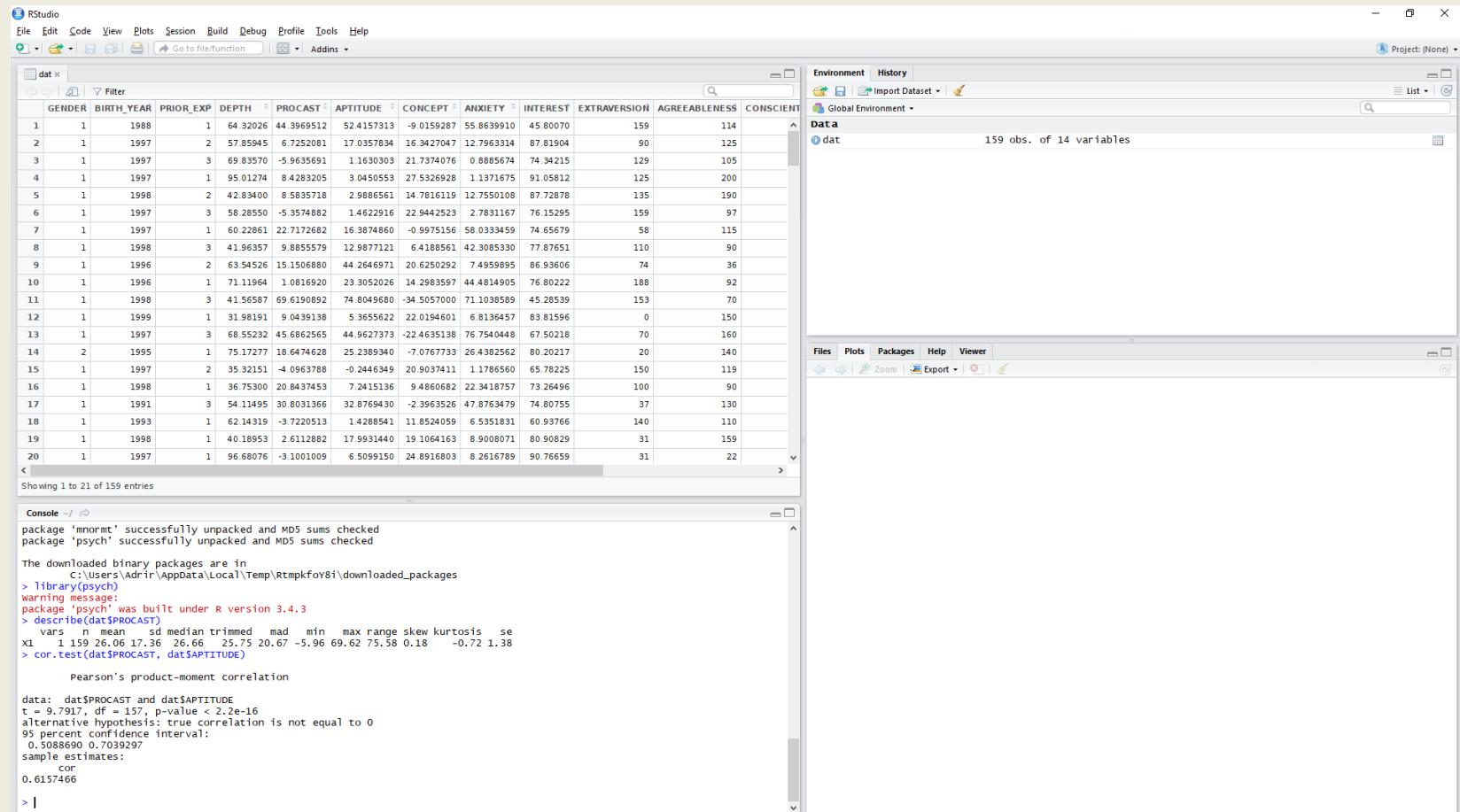
Try:

```
> cor.test(dat$PROCAST, dat$APTITUDE, method="pearson")
```

What does “2.2e-16” mean?

What can we conclude from this data?

Correlation



The screenshot shows an RStudio interface with the following components:

- Data View:** A grid showing a dataset named "dat" with 159 observations and 14 variables. The variables include GENDER, BIRTH_YEAR, PRIOR_EXP, DEPTH, PROCAST, APTITUDE, CONCEPT, ANXIETY, INTEREST, EXTRAVERSION, AGREEABLENESS, and CONSCIENT.
- Environment View:** Shows the "dat" object with the description "159 obs. of 14 variables".
- Console View:** Displays the R code used for the analysis, including package imports, data loading, and correlation calculations.

```

library(psych)
#> library(ROCAST)
#> describe(dat$PROCAST)
#> vars n mean sd median trimmed mad min max range skew kurtosis se
#> X1 1 159 26.06 17.36 26.66 25.75 20.67 -5.96 69.62 75.58 0.18 -0.72 1.38
#> cor.test(dat$PROCAST, dat$APTIITUDE)

Pearson's product-moment correlation

data: dat$PROCAST and dat$APTIITUDE
t = 9.7917, df = 157, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5088690 0.7039297
sample estimates:
cor
0.6157466
  
```

Statistical Tests

- Enable us to assign a confidence value to an observed relationship or difference between groups or treatment conditions
- This is what statisticians call **significance** level (or sometimes **alpha α**)
- Significance is the probability (**p**) that observations as extreme or more extreme were made under the assumption that the **null hypothesis** is true
- A lower significance value implies a lower probability that the result is within expected variance assuming the null hypothesis is true

Statistical tests

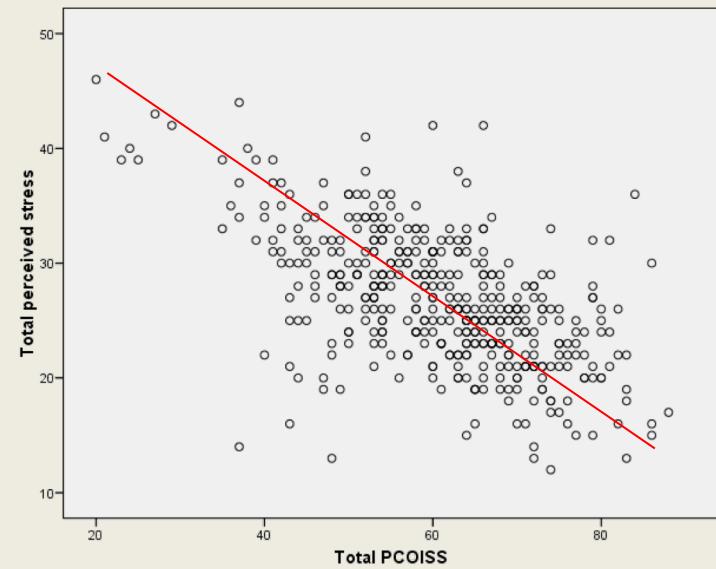
- Significance depends on various factors:
 - Size of difference / degree of relationship
 - Degree of variability or dispersion within the sample(s)
 - Size of the sample
- Conventionally **p = 0.05** is used as the threshold of significance
 - Means 1 in 20 chance that observed difference/relationship is not real
 - Lower values → better (e.g. p = 0.01, p = 0.001)
 - Threshold should be lowered if you are computing many tests on the same data because chance of false positive increases

Linear regression

- Finds a linear model for the relationship between one or more independent variables and a dependent variable
- Between an IV x and a DV y , line of best fit has form

$$y = mx + c$$
- Linear regression finds the coefficients m and c
- Can be used to quantify the relationships of variables, and also for prediction (if correlation is strong)
- Multiple variable linear regression:

$$y = m_1x_1 + m_2x_2 + \dots + c$$



Linear Regression

To regress one independent variable (APTITUDE) against one dependent variable (PROCAST):

```
> rm <- lm(dat$APTITUDE ~ dat$PROCAST)
```

<- is the assignment operator in R

A formula object

To view the analysis use:

```
> summary(rm)
```

What does these results suggest?

Scatter plots

To plot PROCAST (y-axis) against APTITUDE (x-axis):

```
> plot(dat$PROCAST, dat$APTITUDE)
```

To plot a line of best fit (regression line):

```
> abline(rm)
```

Equivalent to:

```
> abline(lm(dat$APTITUDE ~ dat$PROCAST))
```

More on this next week!

Multiple Regression

To regress two independent variables against one dependent variable:

```
> rm <- lm(dat$ANXIETY ~ dat$NEUROTICISM + dat$APTITUDE)
```

To view the analysis use:

```
> summary(rm)
```

A formula object
+ denotes multiple IVs, **not** numerical addition!

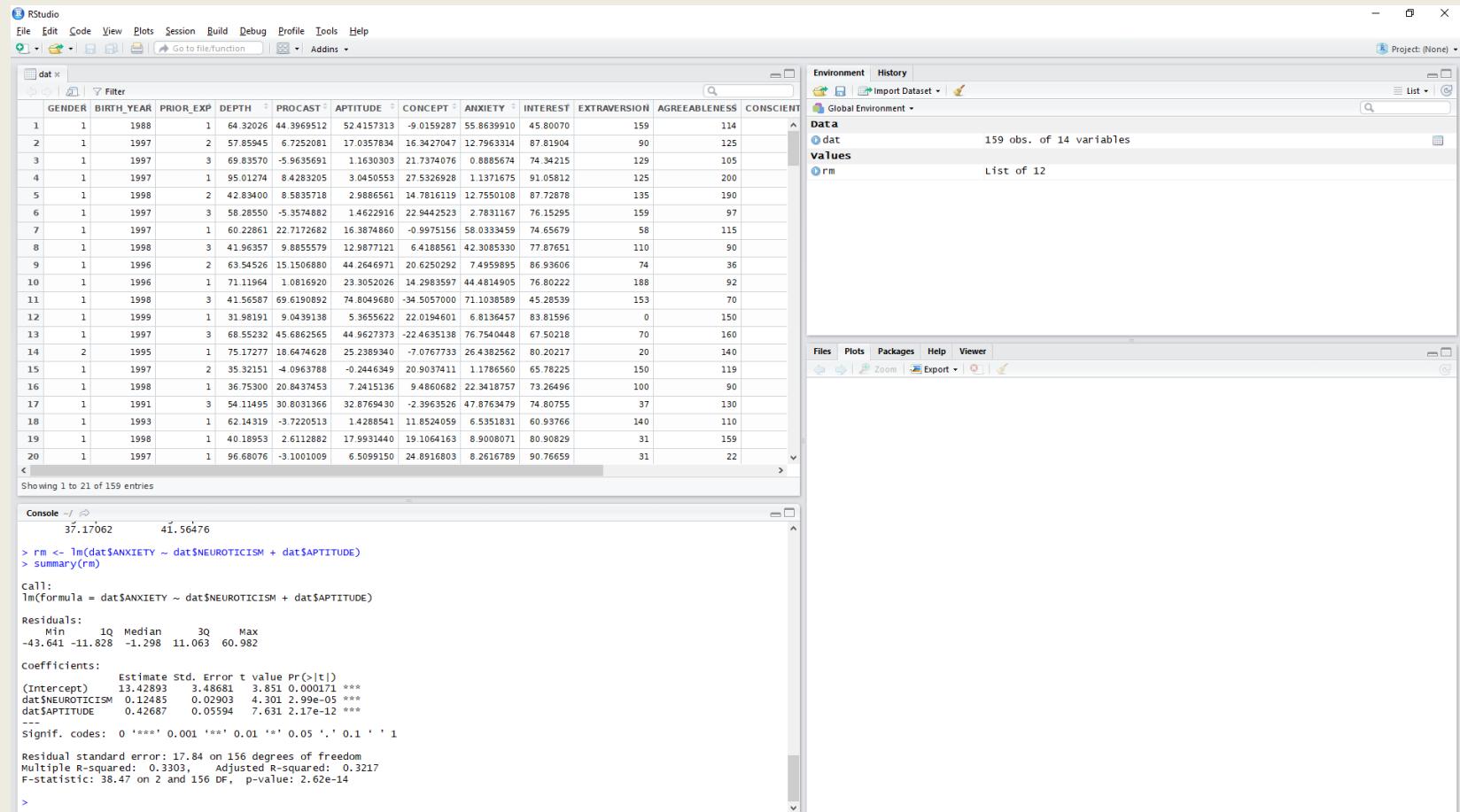
What does these results suggest?

To investigate relative importance of predictors, use:

```
> library(relaimpo)  
> calc.relimp(rm, type=c("pratt"), rela=TRUE)
```

What happens if you wrap this with the plot() function?

Multiple Regression



The screenshot shows an RStudio interface with the following components:

- Environment View:** Shows the global environment with objects `dat` (159 obs. of 14 variables) and `rm` (List of 12).
- Data View:** Displays a data frame named `dat` with 159 rows and 14 columns. The columns include GENDER, BIRTH_YEAR, PRIOR_EXP, DEPTH, PROCAST, APTITUDE, CONCEPT, ANXIETY, INTEREST, EXTRAVERSION, AGREEABLENESS, CONSCIENT, and two unnamed columns. The data is a mix of categorical and numerical values.
- Console View:** Shows the R code used for the analysis, including the creation of a linear model (`lm`) and its summary.

```

Console -> 
37.17062 41.56476
> rm <- lm(dat$ANXIETY ~ dat$NEUROTICISM + dat$APTIITUDE)
> summary(rm)

Call:
lm(formula = dat$ANXIETY ~ dat$NEUROTICISM + dat$APTIITUDE)

Residuals:
    Min      1Q      Median      3Q      Max  
-43.641 -11.828   -1.298   11.063   60.982  

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.42893  3.48681  3.851 0.000171 ***
dat$NEUROTICISM 0.12485  0.02903  4.301 2.99e-05 ***
dat$APTIITUDE  0.42687  0.05594  7.631 2.17e-12 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.84 on 156 degrees of freedom
Multiple R-squared:  0.3303, Adjusted R-squared:  0.3217 
F-statistic: 38.47 on 2 and 156 DF,  p-value: 2.62e-14
  
```

Multiple Regression

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

dat

GENDER BIRTH_YEAR PRIOR_EXP DEPTH PROCAST APTITUDE CONCEPT ANXIETY INTEREST EXTRAVERSION AGREEABLENESS CONSCIENT

1	1	1988	1	64.32026	44.3969512	52.4157313	-9.0159287	55.8639910	45.80070	159	114
2	1	1997	2	57.85945	6.7252081	17.0357834	16.3427047	12.7963314	87.81904	90	125
3	1	1997	3	69.83570	-5.9635691	1.1630303	21.7374076	0.8885674	74.34215	129	105
4	1	1997	1	95.01274	8.4283205	3.0450553	27.5326928	1.1371675	91.05812	125	200
5	1	1998	2	42.83400	8.5835718	2.98086561	14.7016119	12.7550108	87.72878	135	190
6	1	1997	3	58.28550	-5.3574882	1.4622916	22.9442523	2.7831167	76.15295	159	97
7	1	1997	1	60.22861	22.7172682	16.3874860	-0.9975156	58.033459	74.65679	58	115
8	1	1998	3	41.96357	9.8855579	12.9877121	6.4188561	42.3085330	77.87651	110	90
9	1	1996	2	63.54526	15.1506880	44.2646971	20.6250292	7.4959895	86.93606	74	36
10	1	1996	1	71.11964	1.0816920	23.3052026	14.2983597	44.4814905	76.80222	188	92
11	1	1998	3	41.56587	69.6190892	74.8049680	-34.5057000	71.1038589	45.28539	153	70
12	1	1999	1	31.98191	9.0439138	5.3655622	22.0194601	6.8136457	83.81598	0	150
13	1	1997	3	68.55232	45.6862565	44.9627373	-22.4635138	76.7540448	67.50218	70	160
14	2	1995	1	75.17277	18.6474628	25.2389340	-7.0767733	26.4382562	80.20217	20	140
15	1	1997	2	35.32151	-4.0963788	-0.2446349	20.9037411	1.1786560	65.78225	150	119
16	1	1998	1	36.75300	20.8437453	7.2415136	9.4860682	22.3418757	73.26496	100	90
17	1	1991	3	54.11495	30.8031366	32.8769430	-2.3963526	47.8763479	74.80755	37	130
18	1	1993	1	62.14319	-3.7220513	1.4288541	11.8524059	6.5351831	60.93766	140	110
19	1	1998	1	40.18953	2.6112882	17.9931440	19.1064163	8.9008071	80.90829	31	159
20	1	1997	1	96.68076	-3.1001009	6.5099150	24.8916803	8.2616789	90.76659	31	22

Show 1 to 21 of 159 entries

Console

```

am1

Attaching package: 'survey'
The following object is masked from 'package:graphics':
  dotchart

Loading required package: mitools
This is the global version of package relaimpo.

If you are a non-US user, a version with the interesting additional metric pmvd is available
from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.

Warning messages:
1: package 'relaimpo' was built under R version 3.4.3
2: package 'mitools' was built under R version 3.4.3
> plot(calc.relimp(rm,type=c("pratt"), rela=TRUE))
>
> rm <- lm(ANXIETY ~ NEUROTICISM + APTITUDE, data=dat)
> plot(calc.relimp(rm,type=c("pratt"), rela=TRUE))
>
>
```

Environment History

Global Environment

Data dat 159 obs. of 14 variables

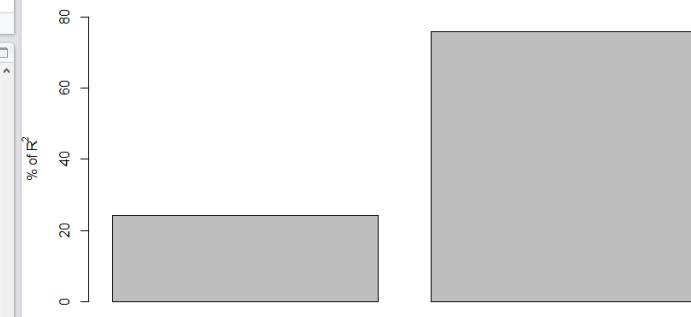
values

rm List of 12

Files Plots Packages Help Viewer

Relative importances for ANXIETY

Method Pratt



% of R²

NEUR

APTI

R² = 33.03%, metrics are normalized to sum 100%.

T-Test

To see a correlation:

```
> t.test(x~y)
```

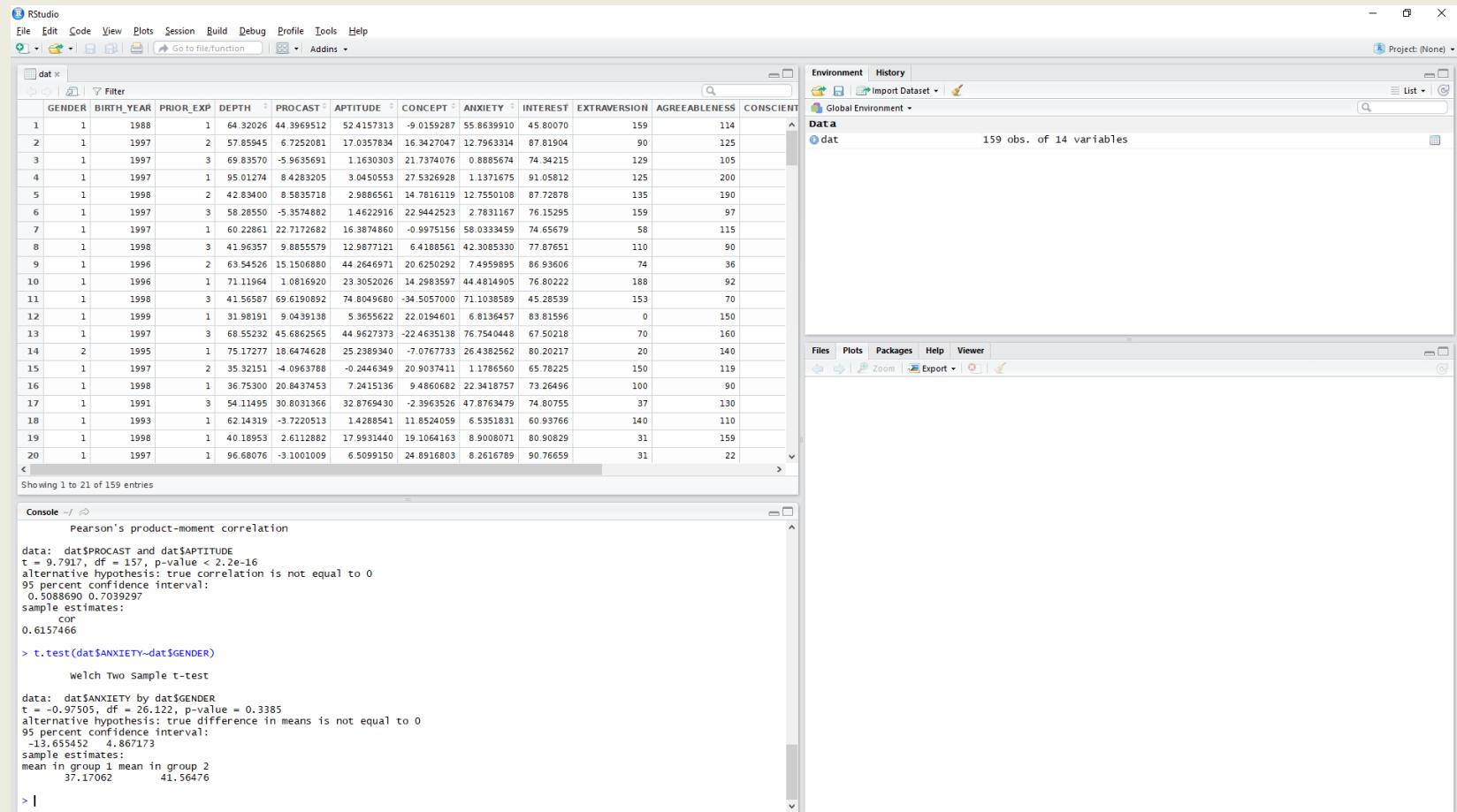
Note: Data must be in long-format. The tilde indicates the grouping variable.

Try:

```
> t.test(dat$ANXIETY~dat$GENDER)
```

What can we conclude from this data?

T-Test



The screenshot shows an RStudio interface with the following components:

- Environment View:** Displays the dataset "dat" with 159 observations and 14 variables.
- Global Environment View:** Shows the same dataset "dat".
- Console View:** Contains the R code and output for the T-Test analysis.

Console Output:

```

Pearson's product-moment correlation

data: dat$PROCAST and dat$APTITUDE
t = 0.7917, df = 157, p-value = 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5088690 0.7039297
sample estimates:
 cor
0.6157466

> t.test(dat$ANXIETY~dat$GENDER)

Welch Two Sample t-test

data: dat$ANXIETY by dat$GENDER
t = -0.97505, df = 26.122, p-value = 0.3385
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.655452 4.867173
sample estimates:
mean in group 1 mean in group 2
 37.17062 41.56476

```

Other Analyses

- There is a whole host of analyses that can be conducted in Rstudio!
- Investigate for yourself those which will have relevance to your particular project.
- For resources, see:
 - www.rdocumentation.org
 - www.statmethods.net
 - <http://tutorials.iq.harvard.edu/R/>
 - <https://support.rstudio.com/hc/en-us>