COMP360

RESEARCH DISSERTATION

Data Analysis and Inference with Statistical Tests

Objectives for Today

After this session, you should be able to:

- **Explain** the link between making inferences from research data and statistical analysis
- **Recall** key descriptive statistics such as measures of central tendencies and dispersion
- **Explain** the foundations of null-hypothesis significance testing

Objectives for Today

After this session, you should be able to:

- **Outline** the importance of validity **and** reliability in data analysis
- **Suggest** the **most** appropriate statistical test to apply to given data
- **Configure** R for statistical data analysis
- **Interpret and explain** inferences that can be made from the output of statistical tests conducted in R

GAMES ACADEMY | FALMOUTH UNIVERSITY

Objectives for Today

1. Using Statistical Tests to Support Inference

- Foundations of Quantitative Data Analysis for Inference
- Discrete and Continuous Data

2. Research Quality

- Validity and Reliability
- False Discovery and Replicability

3. Common Statistical Tests

- Measures of Relation versus Measures of Difference
- Correlations, Chi-Square, and Regression
- F-tests, t-tests and ANOVA
- Interpreting p-values from correlations

Objectives for Today

4. Importing Data into R

- Importing a .csv file
- Checking the data with view() and summary()

5. Data Analysis in R

- Correlational Analysis
- t-test
- Linear Model Multiple Regression

6. Further Exploration of R

- Validity and Reliability
- False Discovery and Replicability

Using Statistical Tests to Support Inference

Making Claims Beyond Your Observations

Statistical Inference

- Why analyse quantitative data?
- Aim to generalise the findings from researcher's observations to a broader population
  - Known as 'inference'
  - Derive a logical conclusion from premises known to be true
  - Applying statistical analyses to data from a survey or experiment
- Ecological fallacy
  - Inferences about specific individuals **cannot be made** based solely on statistics collected for the group to which those individuals belong

Statistical Inference

> *Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is 'How sure are we that the estimated mean is near the true population mean?'. While the equations and details change depending on the research context, the foundations for inference are the same.*

(Diez, Barr & Cetinkaya-Rudel, 2015, p. 168)

Quantitative Data Analysis

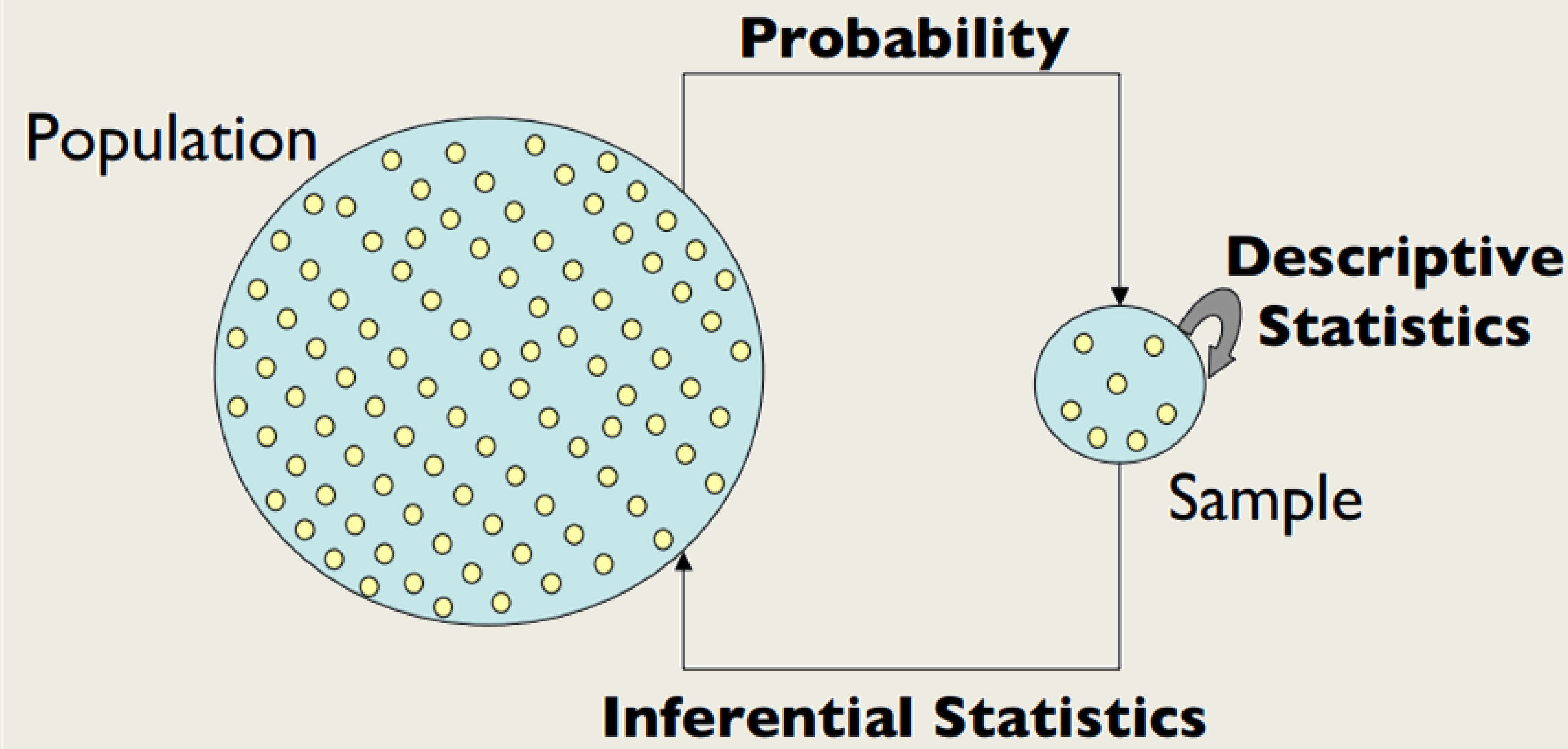
- **Population**
  A set of items that share a characteristic; a total set from which observations may be drawn.
  - Some populations are very large
    - Possible combinations of a deck of cards:
      ~50 quintillion
    - People:
      ~7 billion
    - “British”
      65.6 million
    - Falmouth Students
      5,446
  - Too many to conduct a census to involve everyone in the research!

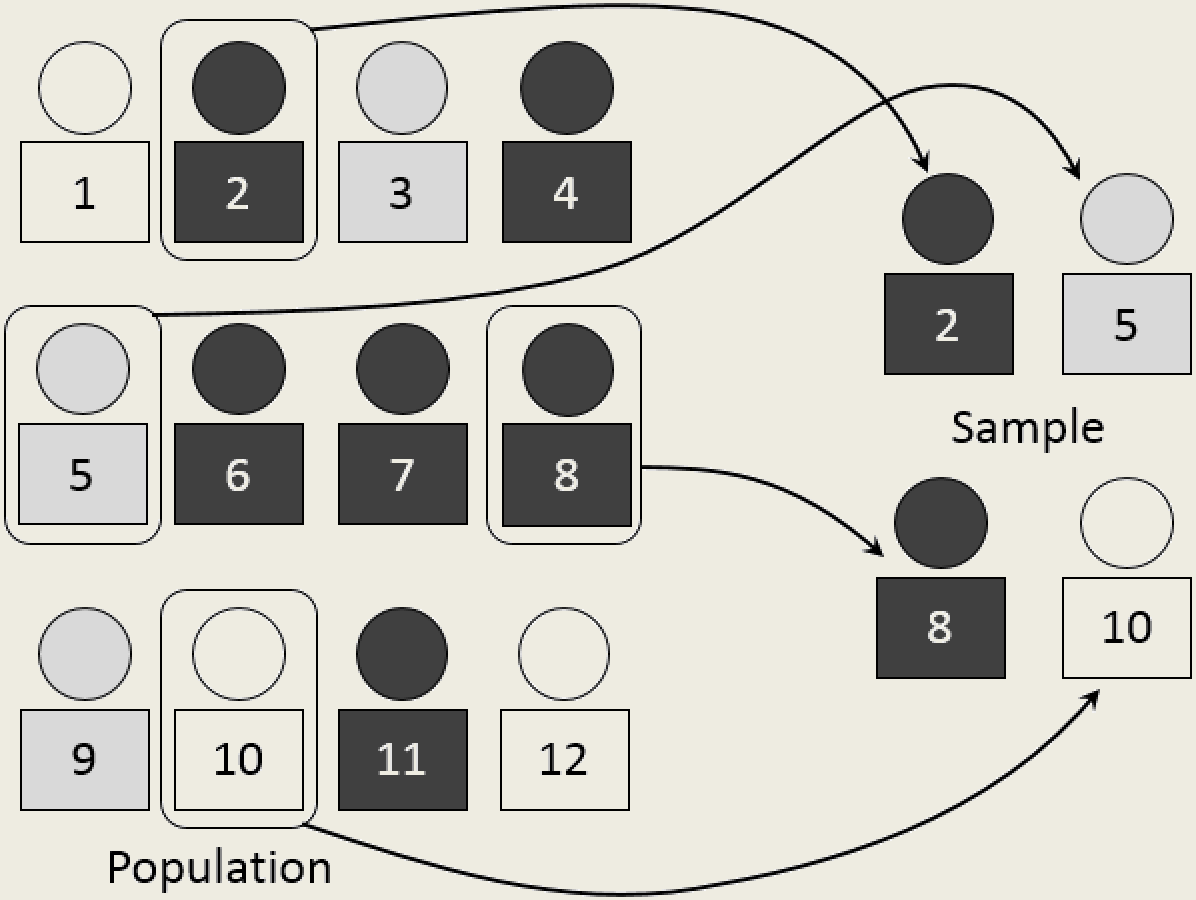
Quantitative Data Analysis

- **Sample**

  A set of observations of items collected and/or selected from a population by a defined procedure.

  Researchers collect data from a *sample* of items belonging to a broader *population* they are interested in
  - Ideally, randomly
  - Consider "representativeness"


Probability
Population
Descriptive
Statistics
Sample
Inferential Statistics


1
2
3
4
5
6
7
8
9
10
11
12
Population
2
5
Sample
8
10

Sampling Methods

Probability

- Random
- Stratified Random
- Systematic

Non-Probability

- Convenience
- Quota
- Purposive
- Snowball

Exploratory Data Analysis

- Before making inferences from data it is essential to examine all your variables.
- Why listen to the data?
  - Catch mistakes
  - See patterns in the data
  - Find violations of assumptions
  - Generate new hypotheses
- …and because if you don't, you will have trouble later

Discrete vs Continuous Data

A set of data is said to be **continuous** if the values belonging to the set can take on **any** value within a finite or infinite interval.

Analysts typically measure values using instruments, and arrive at interval or ratio values.

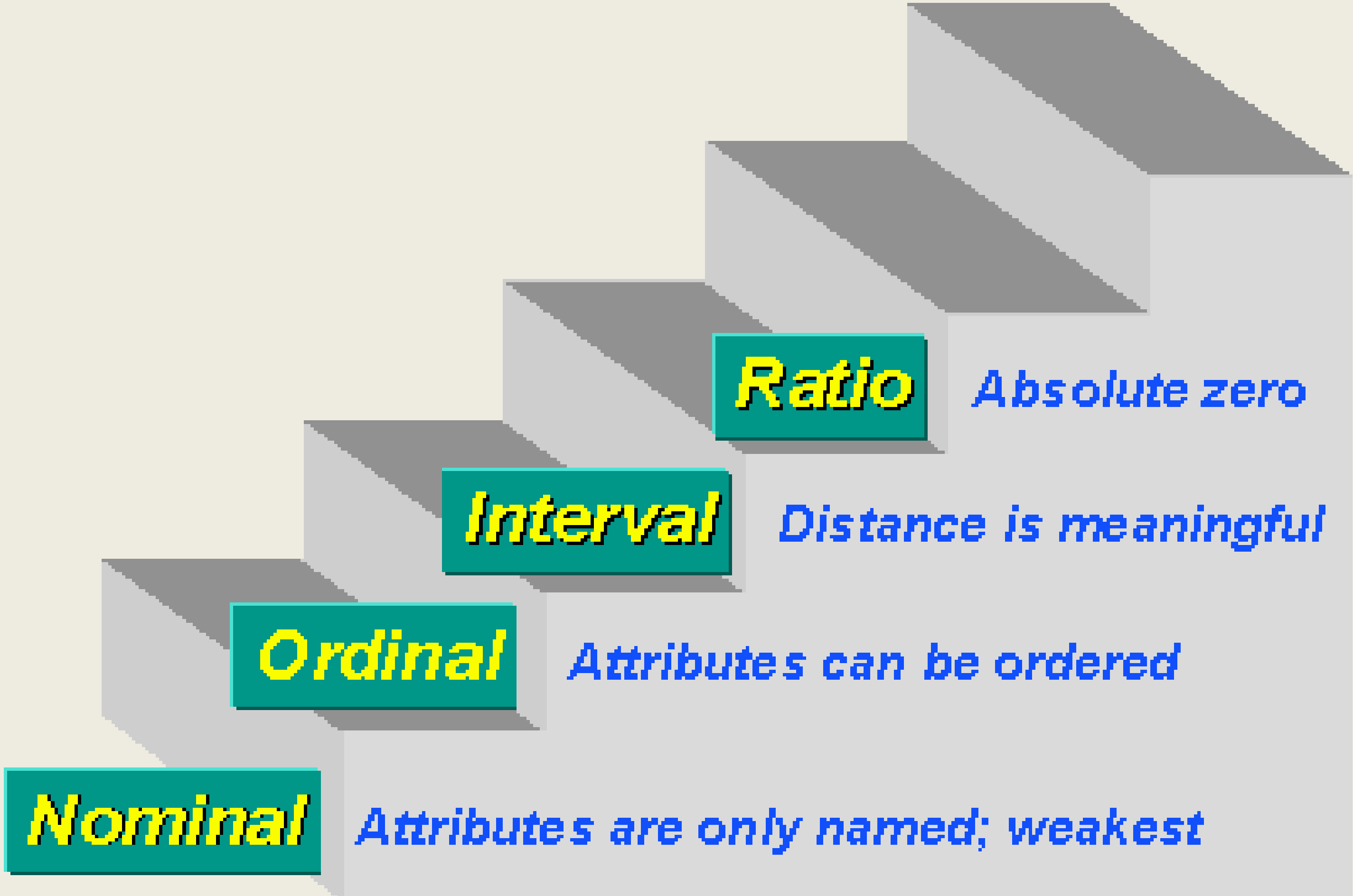
Examples: successful hits (interval); velocity (ratio).

A set of data is said to be **discrete** if the values belonging to the set are **distinct** and **separate**.

Analysts typically count the occurrences of nominal or ordinal values in datasets for analysis.

Examples: ethnicity (nominal); age-bracket (ordinal).

Levels of Measurement



Analysing Discrete Data

- Discrete variables are those which have a defined set of distinct values; often, these are categories:
  - Gender, nationality, marital status, grade, education level
- Level of measurement?
  - Nominal
  - Ordinal
- It is usually valuable to know **how many cases** belong to each group as defined by your categorical variables
  - *How many cases are there in each category group?*
- Most statistical tests require
  - roughly equal numbers of cases in each group
  - at least 4 cases to estimate t (but more = better) in all groups
  - Strive for low standard error
- ...in order to produce **dependable** results

Analysing Discrete Data

- Example: *"Are there national differences in the ease of learning particular game genres?"*
- What was the relative proportion of British and non-British observations in your sample?
  - Demographic information is normally reported in the method section of your report
  - Is the sample suitably representative of the population of interest?
  - Is there enough data in each group to make a meaningful comparison?
- You may need to collect more data
  - If the sample does not reflect the demographic structure of the target population
  - If the populations under study radically unbalanced or particular cells in factorial analyses are too small

Analysing Discrete Data

- In *R* to obtain descriptive statistics for categorical variables, select:
  - Sapply()
  - Summary()
  - Describe()

Statistics

gender

N	Valid	271
	Missing	0

gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	84	31.0	31.0	31.0
	male	187	69.0	69.0	100.0
	Total	271	100.0	100.0	

- Comparing the groups may be problematic in this case. Why?

Analysing Discrete Data

- Frequency tables may also be sufficient to fully answer simple questions
- e.g. Running a count on responses to a question:
  - “Which is your favourite game genre?”
- Could also provide a rank order of popularity
- But no test of statistical ‘significance’
  - For example: in a sample of 200, 99 male respondents and 101 female respondents said that they play Overwatch.
  - So, can we conclude that there are more *female players than male players on Overwatch*?

Analysing Discrete Data

- Frequency tables may also be sufficient to fully answer simple questions
- e.g. Running a count on responses to a question:
  - "Which is your favourite game genre?"
- Could also provide a rank order of popularity
- But no test of statistical 'significance'
  - For example: in a sample of 200, 99 male respondents and 101 female respondents said that they play Overwatch.
  - So, can we conclude that there are more *female players than male players on Overwatch*? **No!**

Analysing Continuous Data

- Continuous variables are scale level measures
  - e.g. Age, weight, height, time, test scores, number of customers.
- Level of measurement?
  - Interval
  - Ratio
- Makes little sense to look at frequency tables, though bucketing and histograms can provide insight into likely distribution of values
- More interested in **summary statistics** such as:
  - Measures of central tendency
  - Measures of variability (i.e., dispersion)
  - Measures of normality

Analysing Continuous Data – Central Tendency

The central tendency is the value which all the data tends towards:

- **Mean**
  - Sum of all values divided by the number of values (N or n)
  - Most statistical difference tests are based on this
- **Median**
  - Arrange values in order and pick the middle value (if N is odd) or average the two middle numbers (if N is even)
  - More useful for ordinal and non-normal data
  - Unaffected by extreme scores
- **Mode**
  - Most frequent value in the data set
  - Not at all interesting for continuous data
  - More useful for categorical data (nominal and ordinal)

GAMES ACADEMY | FALMOUTH UNIVERSITY

Analysing Continuous Data – Central Tendency

The following numbers represent the seconds (in minutes) that players spent on a loading screen for a new level are: **4, 5, 1, 3, 24, 5**

Mean? Median? Mode?

- The mean time is: (4+5+1+3+24+5)/6 = 42/6 = **7** seconds
- The median is: 4, 5, 1, 3, 22 →1, 3, 4, 5, 5, 22 → **4.5** seconds
- The mode is: **5** minutes

- What is the median for: **4, 5, 1, 3, 22?** → **4**
- What is the mode for: **4, 5, 1, 3, 22?** → No mode
- What is the mode for: **4, 5, 1, 3, 22, 5, 1?** → **1** and **5** (Two modes)

Analysing Continuous Data – Variability

- Reporting the central tendency is not enough to describe your data
- We also need to know how much the data values vary (i.e. how spread out they are).

- Example: two groups, each with 10 people, self-report the number of hours they play online daily.

- Group 1 data: **4+4+4+4+4+6+6+6+6+6**

- Group 2 data: **0+0+0+0+0+10+10+10+10+10**

- The mean of both datasets is **5** hours.
- So the data is similar...? **Not really!**

Analysing Continuous Data – Variability

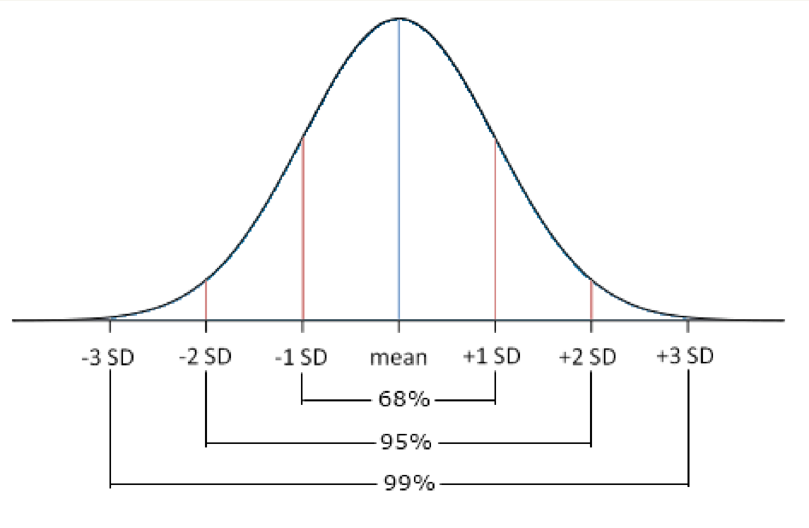
- Standard deviation (SD)
  - Average difference of values from the sample mean
  - Larger standard deviation indicates greater differences between the scores
- In the previous slide (Group 1 SD = 1 vs. Group 2 SD = 5)
- In a normal distribution ~68% of values lie within 1 SD of the mean
- Key statistic in difference tests

Analysing Continuous Data – Variability

- Range
  - Difference between lowest score and highest score
  - *Simple way of identifying **outliers** (extreme values), but doesn't say anything about the variability within the dataset.*
  - *Often used to define sample ages in the "Method" section of a report*

Analysing Continuous Data – Normality

- Height, weight, IQ and other physical measures follow a **normal (bell shaped curve) distribution** of values
- For a large enough population sample, many observations will result in a normal distribution
- Many statistical tests assume that data within a variable is distributed normally
- SD can be used to determine the proportion of values that lie within a particular range.



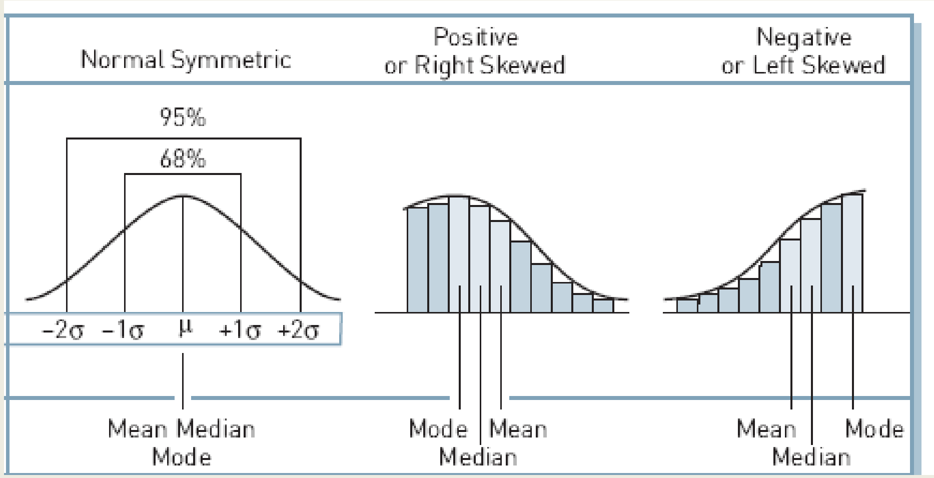
Analysing Continuous Data – Normality

1. **Kurtosis**
   - The ‘peakedness’ of the distribution
   - Zero kurtosis score means normal
   - **Negative** kurtosis → distribution is flat; more extreme cases than expected
   - **Positive** kurtosis → ‘pointy’ distribution; values tend to cluster around centre
   - Excessive kurtosis are normally **only a problem for small samples (N < 200)**

Analysing Continuous Data – Normality

2. Skewness

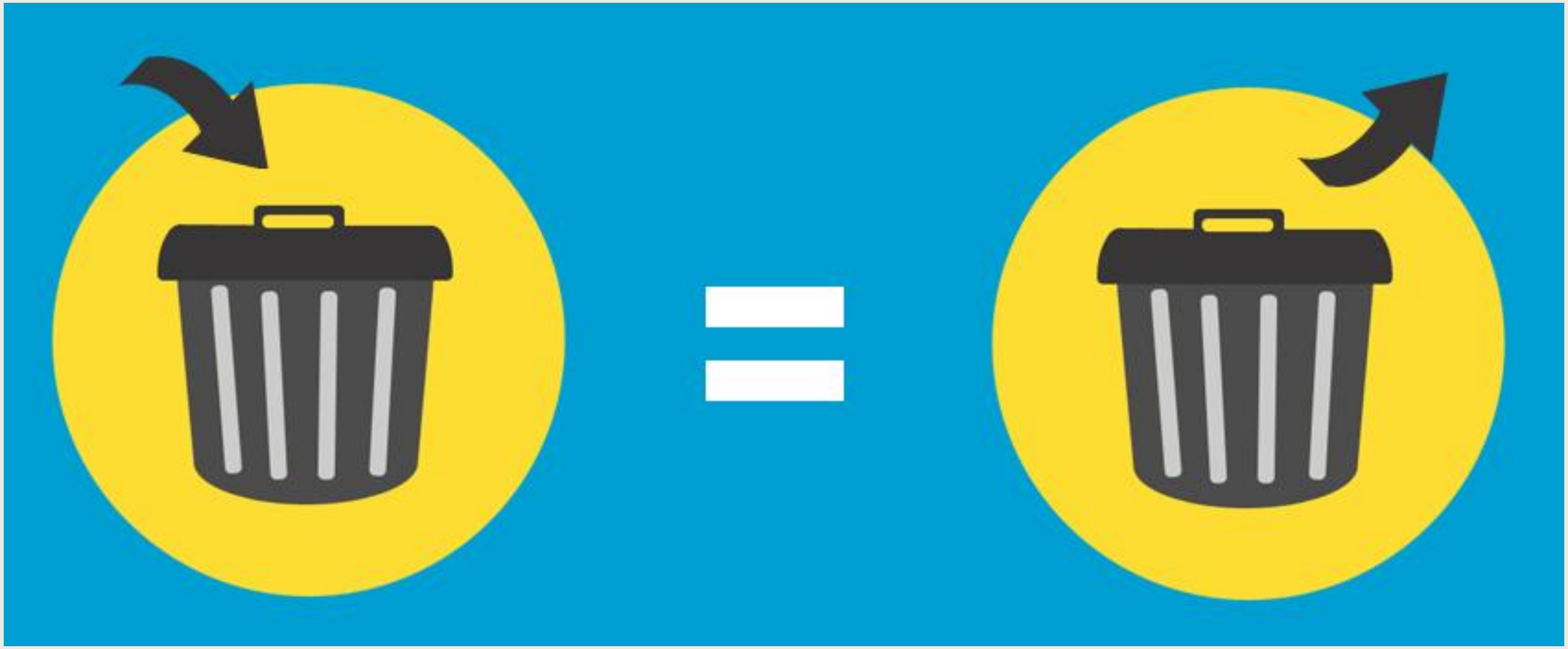
- The symmetry of the distribution
- **Positive** skew → more low values than expected
- **Negative** skew → more high values than expected
- Example: Student scores in a too hard/too easy exam



30

Statistics & Quality

Garbage In, Garbage Out



Garbage In, Garbage Out

Research Quality

- All research is subject to limitations
- However, not all research is equal
    - Unique resources or contexts available
    - “Natural” experimental conditions
    - Funding
- The quality of a research finding is often called its *rigour* and relates to how inherent limitations have been overcome by the researcher

Research Quality

- There are three main components to research quality:
    - Validity
        - Measurement Validity (e.g. Construct Validity)
        - Internal Validity
        - External Validity
    - Reliability
    - Replicability

Validity

Construct validity means selecting the most appropriate measurement tool for the concepts being studied.

Does your tool really measure what you want to assess?

Validity

Internal validity is another term for using different methodological tools to 'triangulate' the data.

What other methods can you use to check for the same phenomenon? Are there flaws in the design of the study that allow for alternative explanations?

Validity

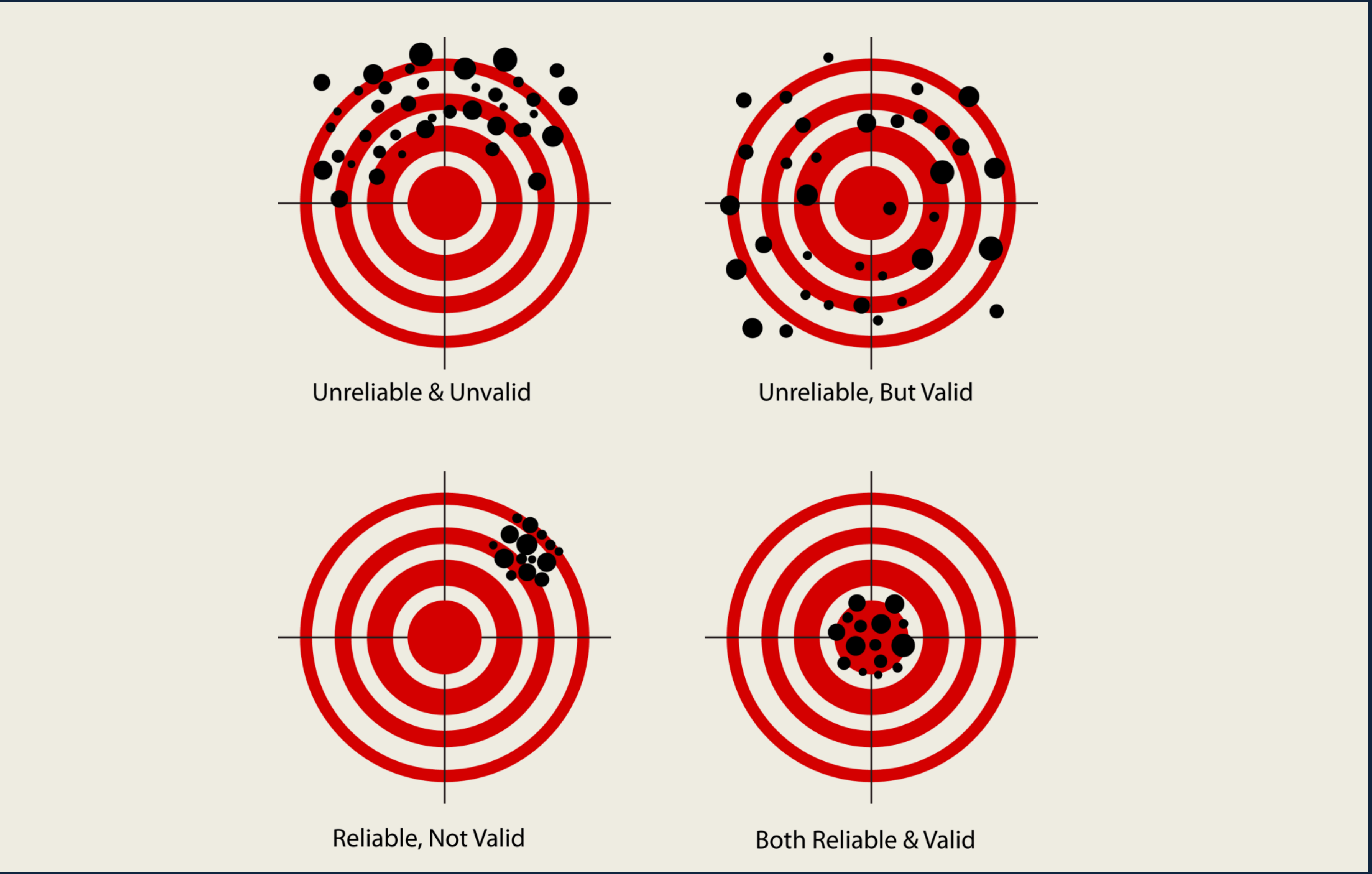
External validity refers to how well the data can be applied beyond the circumstances of the case to more general situations.

Can you apply your data across the industry and to others?

Reliability

Reliability means the extent to which the data collection can be repeated in ways that yield the same data. That is, the extent to which the data itself is accurate.

Are you confident that your study can be repeated by others and the results will be the same?


Unreliable & Unvalid
Unreliable, But Valid
Reliable, Not Valid
Both Reliable & Valid

GAMES ACADEMY | FALMOUTH UNIVERSITY

False Discovery & Replicability

Replicability means the extent to which the results can be reproduced by another researcher working in a similar context. That is, the extent to which the results are accurate and stable.

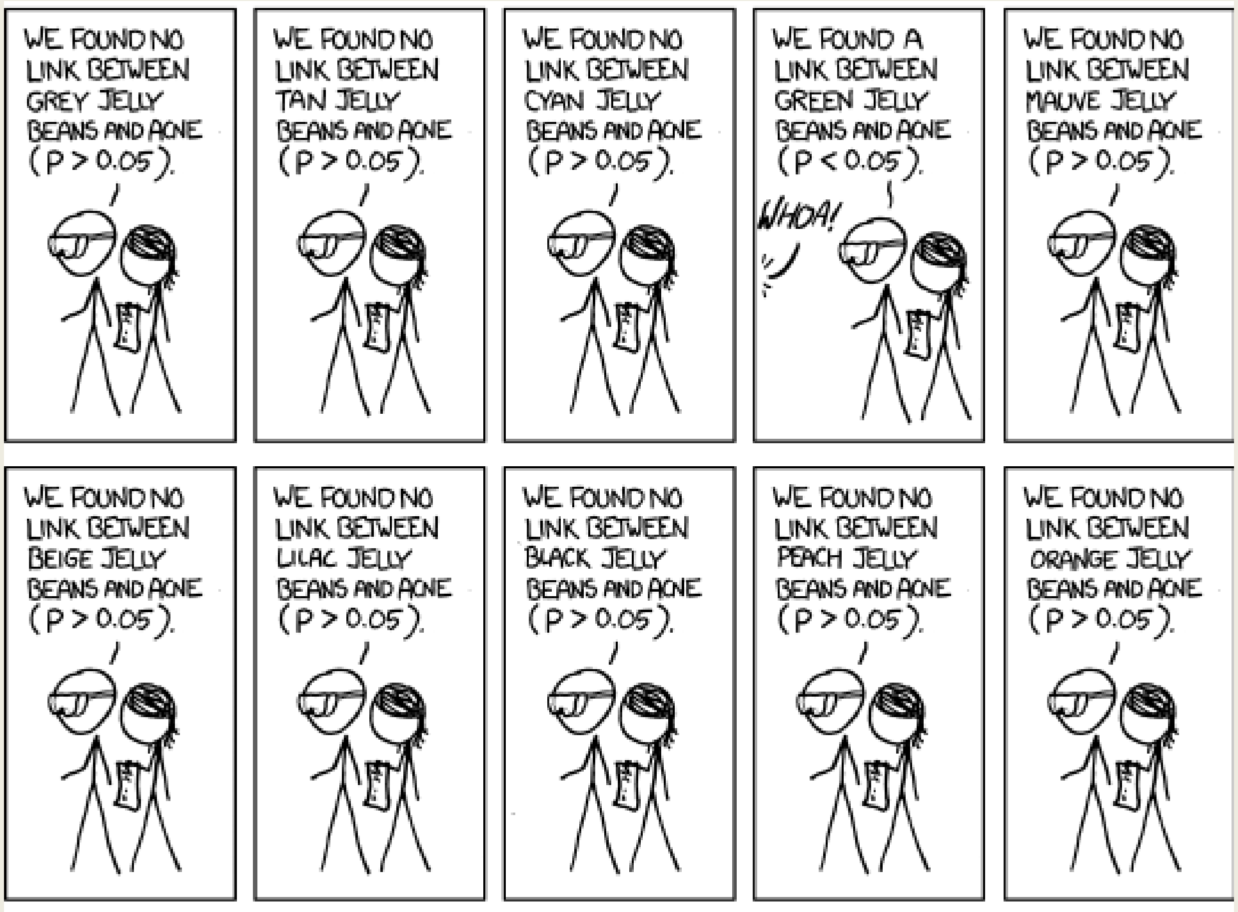
Is there sufficient description to recreate the conditions of the study? Is there a possibility of a Type-I or Type-II error?

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 **levels of stress** and **talent retention**?*
  - *Does **age** predict the **number of loot boxes** purchased?*
  - *Can we model the **level of introversion** of a player based on the number of **in-game messages** they post?*

Research Questions

- Or if there is a **difference** between one or more scores/conditions/groups
  - *Did **weapon** or **armour** skins make the most profit over the last six months?*
  - *Does **working in pairs** improve programming performance?*
  - *Has allowing programmers to **work from home** decreased 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**?

- *'On a scale from 1 to 8, how often do you play games?'*
- Great! I found a difference!!

life sat

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

GAMES ACADEMY | FALMOUTH UNIVERSITY

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 **levels of stress** and **talent retention**?*
- ...will require quite different statistical tests to answer:
    - *Did **weapon** or **armour** skins make the most profit over the last six months?*
    - *Has allowing programmers to **work from home** decrease game development **productivity**?*

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 matricies
  - **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.

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

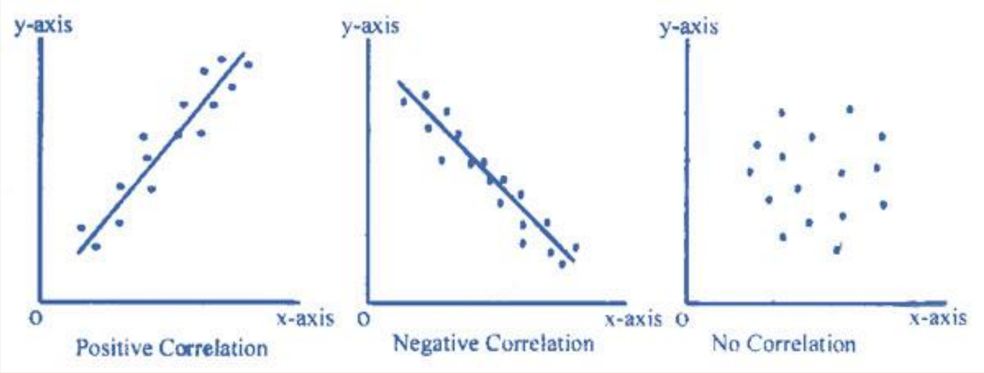
- What is the level of measurement for each variable?
  - Categorical or continuous?
  - Examples of categorical variables?
  - Examples of continuous variables?

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)

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?

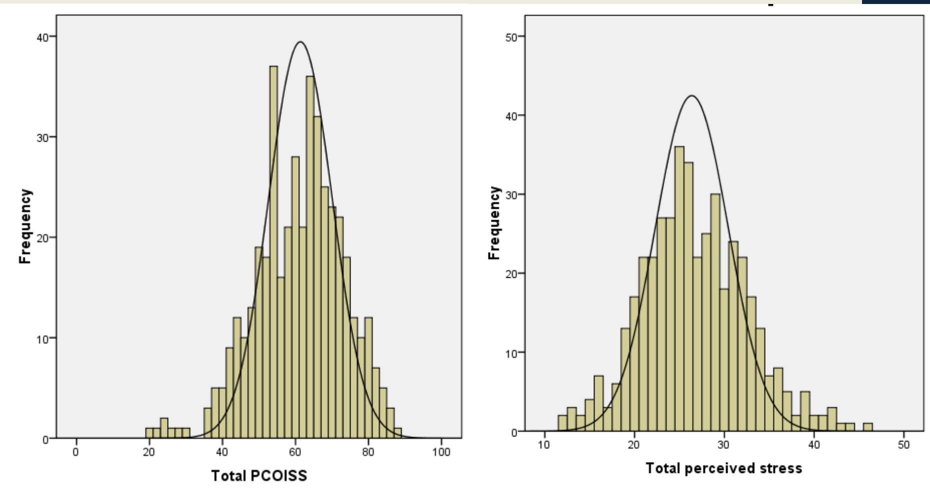


Pearson’s correlation

- 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

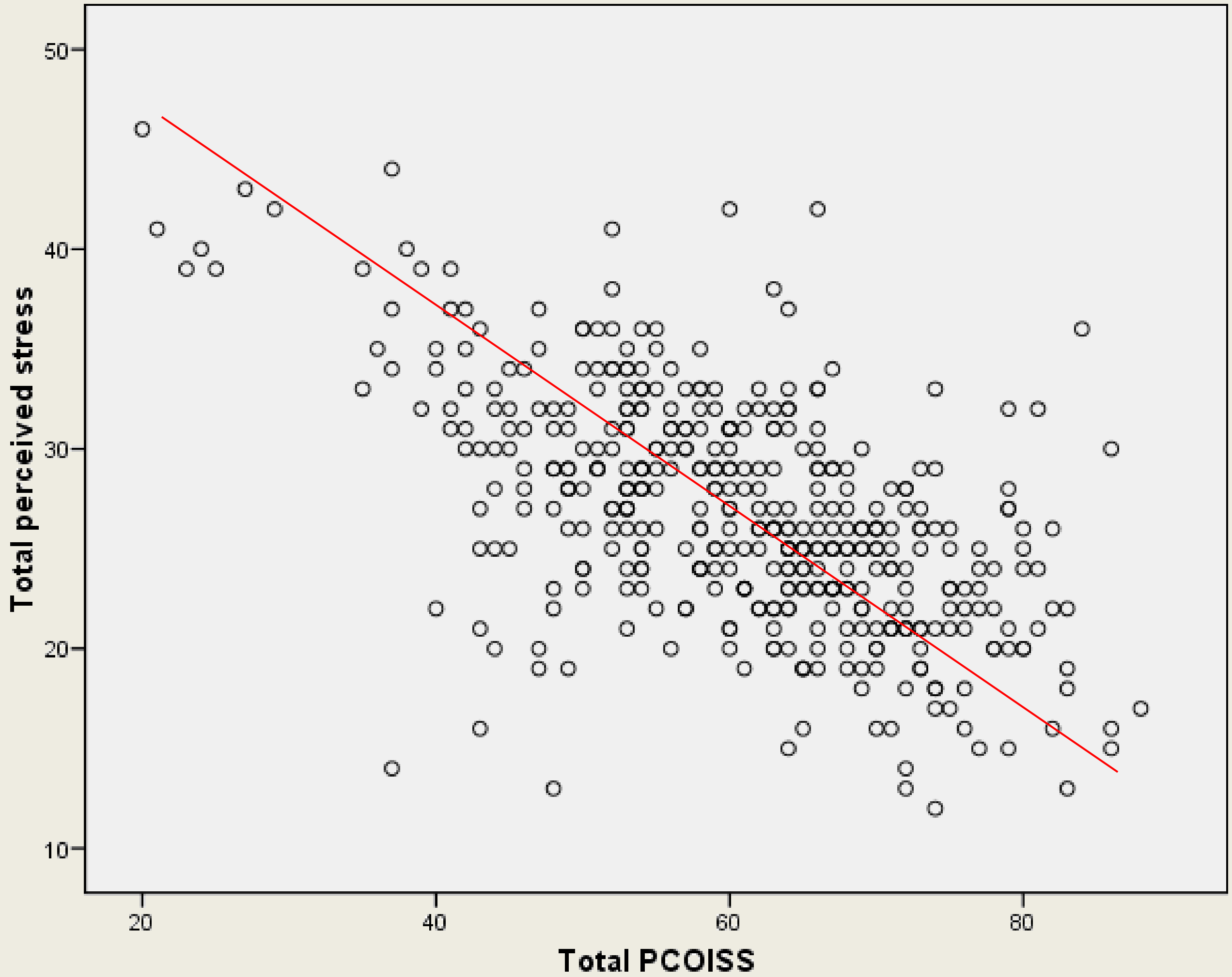
Example: control and stress

- Using Survey data
  - Survey.sav is on the portal
- **Do people with high levels of perceived control (IV) experience lower levels of perceived stress (DV)?**
  - IV = “tpcoiss” (perceived control)
  - DV = “tpstress” (levels of stress)
- Are the data for both normally distributed?



Example

- Looking at the **scatter-plot**
- Positive or negative?
- Strong or weak?



https://campus.datacamp.com/courses/data-visualization-in-r/a-quick-introduction-to-base-r-graphics?ex=3#skiponboarding

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

Computing Pearson r

- Effect Size – some value between -1 and +1
- Significance (p) value - **are we looking for higher or lower values?**
- N – number of valid cases

Correlations

		Total PCOISS	Total perceived stress
Total PCOISS	Pearson Correlation	1	-.581**
	Sig. (2-tailed)		.000
	N	430	426
Total perceived stress	Pearson Correlation	-.581**	1
	Sig. (2-tailed)	.000	
	N	426	433

**. Correlation is significant at the 0.01 level (2-tailed).

Interpreting output - I

- The magnitude of the coefficient is the measure of the strength of relationship
  - Small r = 0.10 to 0.29
  - Medium r = 0.30 to 0.49
  - Large r = 0.50 to 1.0
- It's polarity indicates direction
  - Positive sign means X and Y vary in the same direction
  - Negative sign means X and Y vary in opposite directions
  - r=0.5 is just as strong as r=-0.5; just different direction

Correlations

		Total PCOISS	Total perceived stress
Total PCOISS	Pearson Correlation	1	-.581**
	Sig. (2-tailed)		.000
	N	430	426
Total perceived stress	Pearson Correlation	-.581**	1
	Sig. (2-tailed)	.000	
	N	426	433

**. Correlation is significant at the 0.01 level (2-tailed).

Interpreting output - II

- Significance
  - Remember a p-value of 0.05 or lower is normally deemed significant
  - This only refers to the reliability of the result, not the strength of the relationship (the coefficient tells us that)
  - Even quite small r can be highly significant when N is large

Correlations

		Total PCOISS	Total perceived stress
Total PCOISS	Pearson Correlation	1	-.581**
	Sig. (2-tailed)		.000
	N	430	426
Total perceived stress	Pearson Correlation	-.581**	1
	Sig. (2-tailed)	.000	
	N	426	433

**. Correlation is significant at the 0.01 level (2-tailed).

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")
```

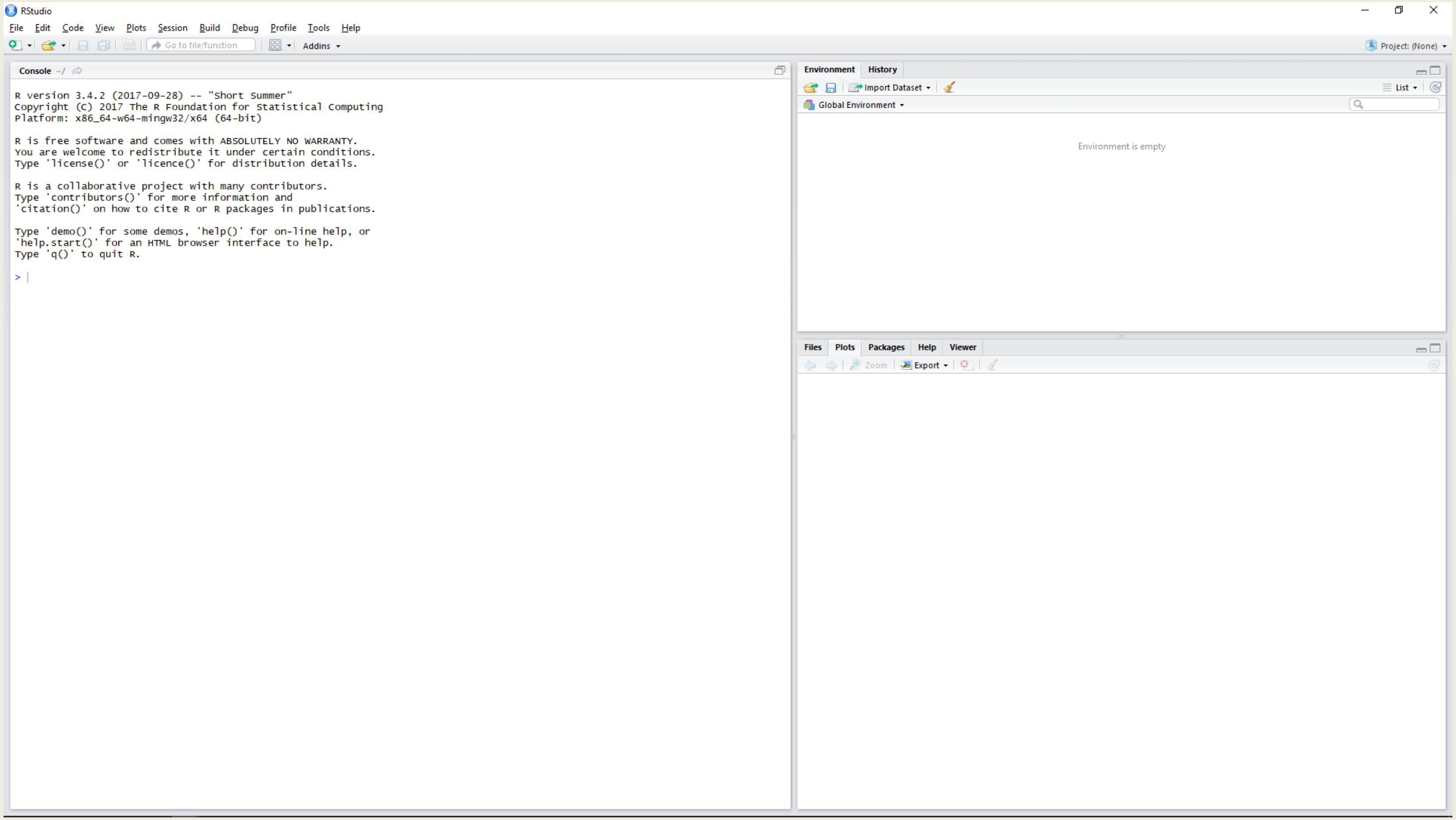
To view the data use:

```
> view(dat)
```

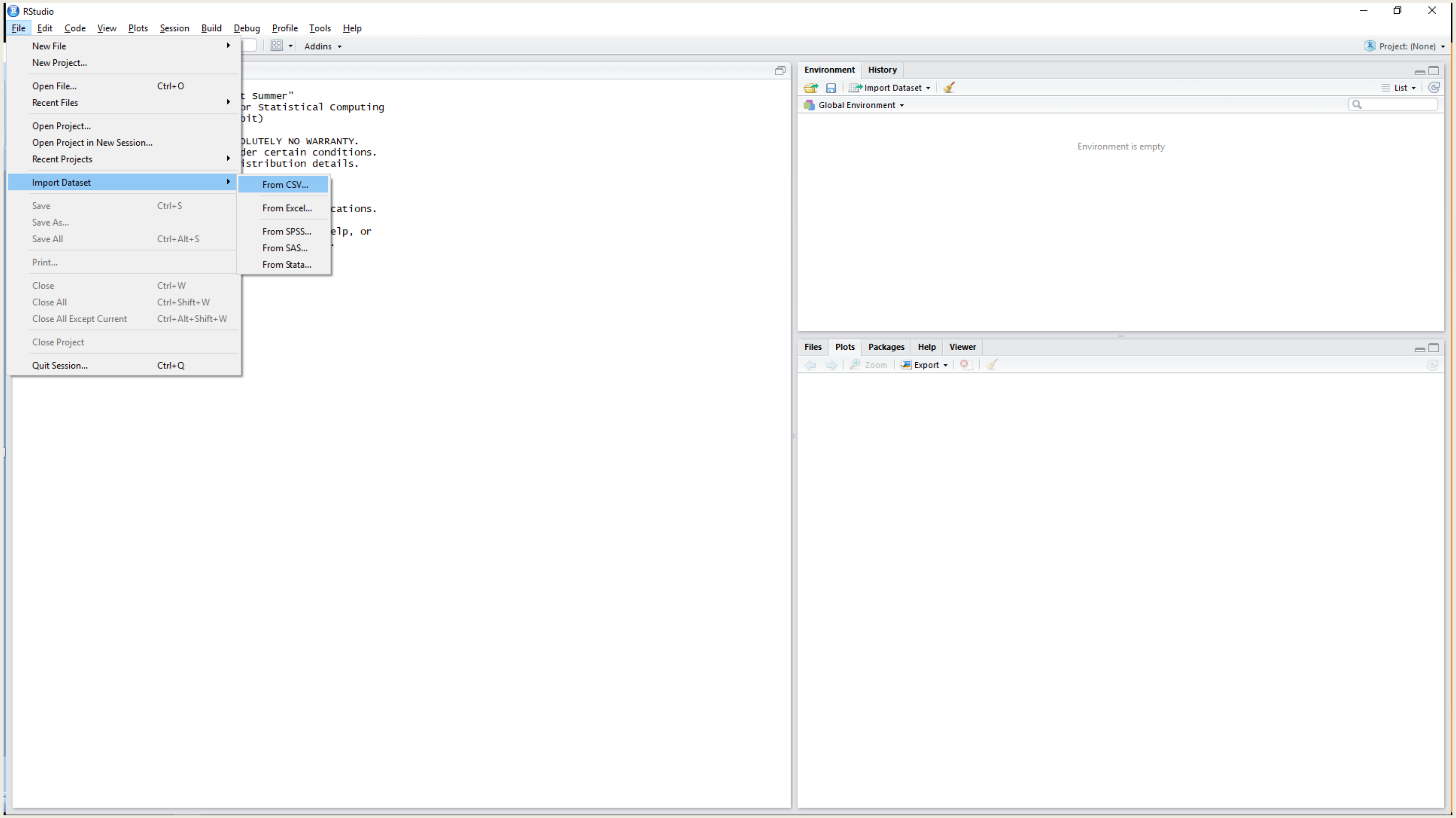
To see the descriptive statistics for all of the variables in the data use:

```
> summary(dat)
```

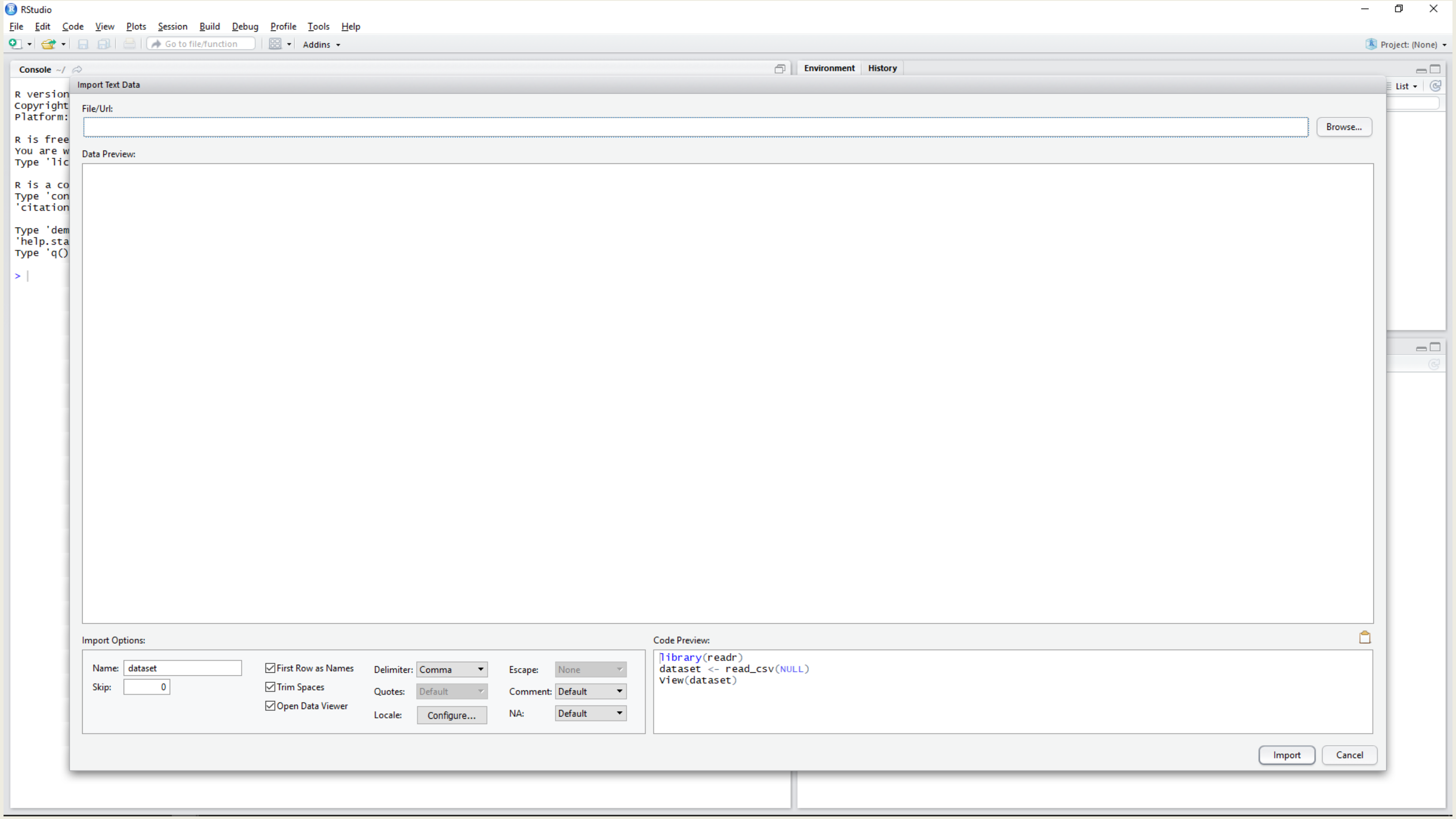
Import Data into R



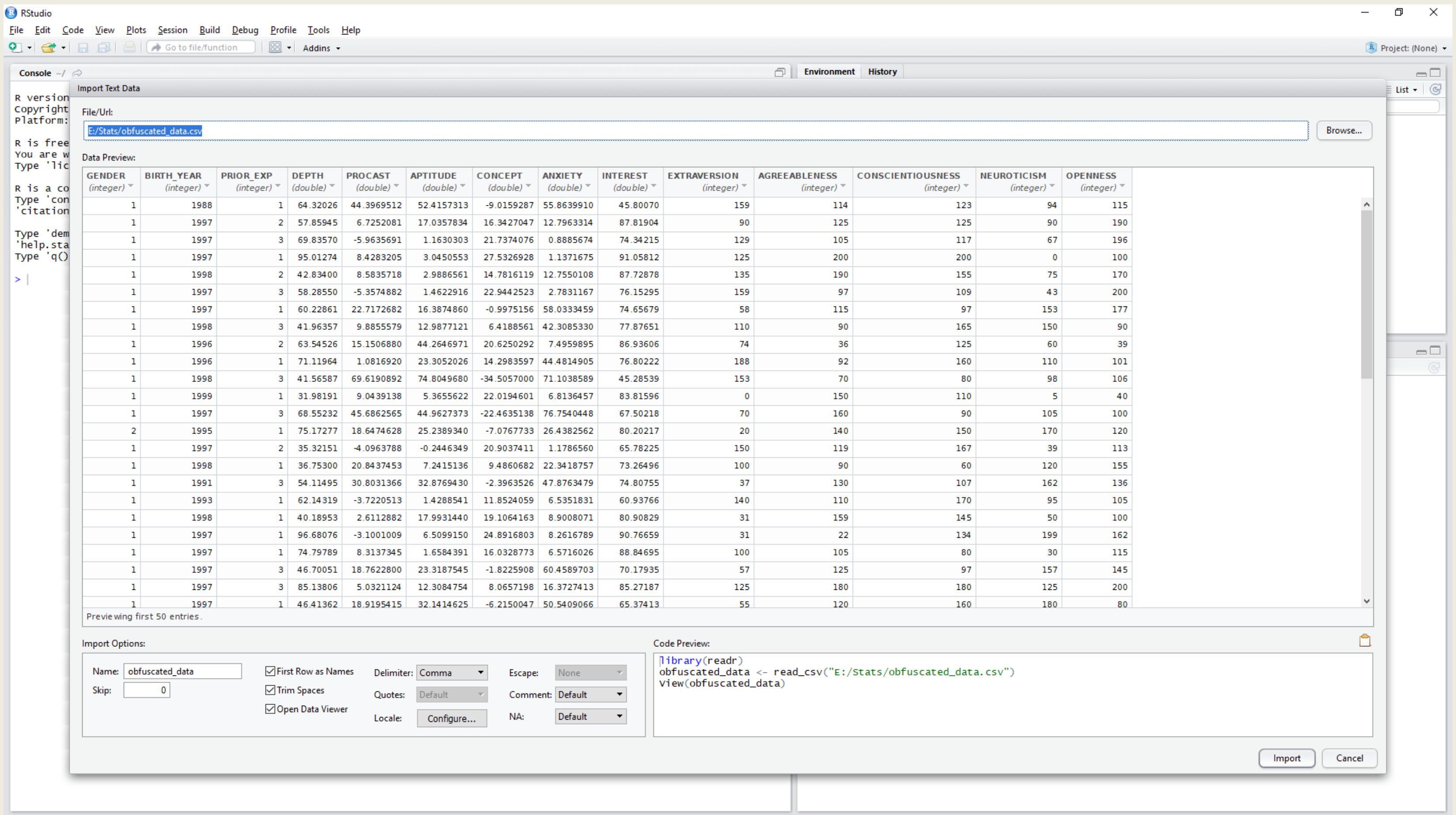
Import Data into R



Import Data into R



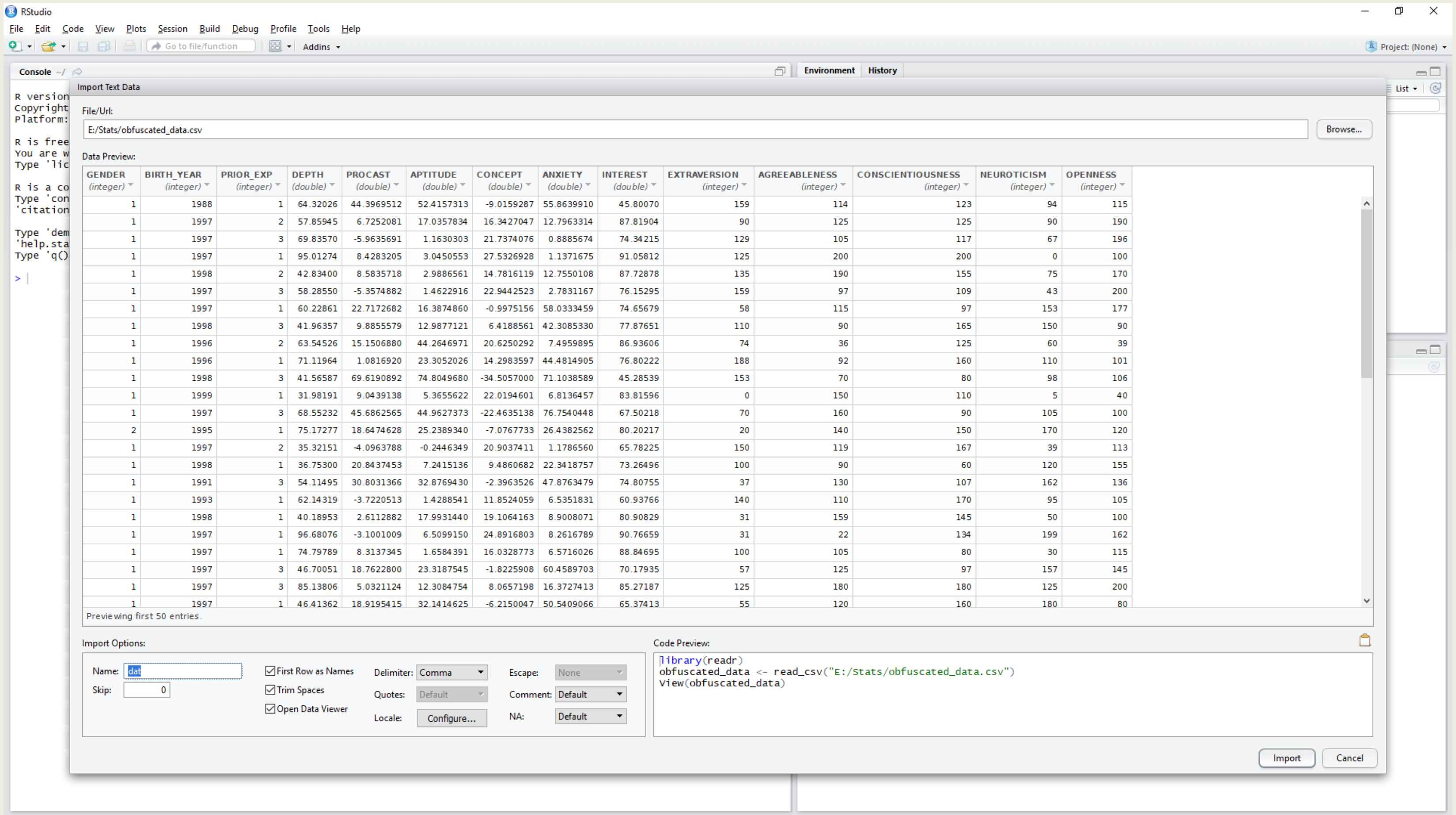
Import Data into R



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	8.4283205	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	58.0333459	74.65679	58	115	97	153	177
1	1998	3	41.96357	9.8855579	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	83.81596	0	150	110	5	40
1	1997	3	68.55232	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	5.0321124	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

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

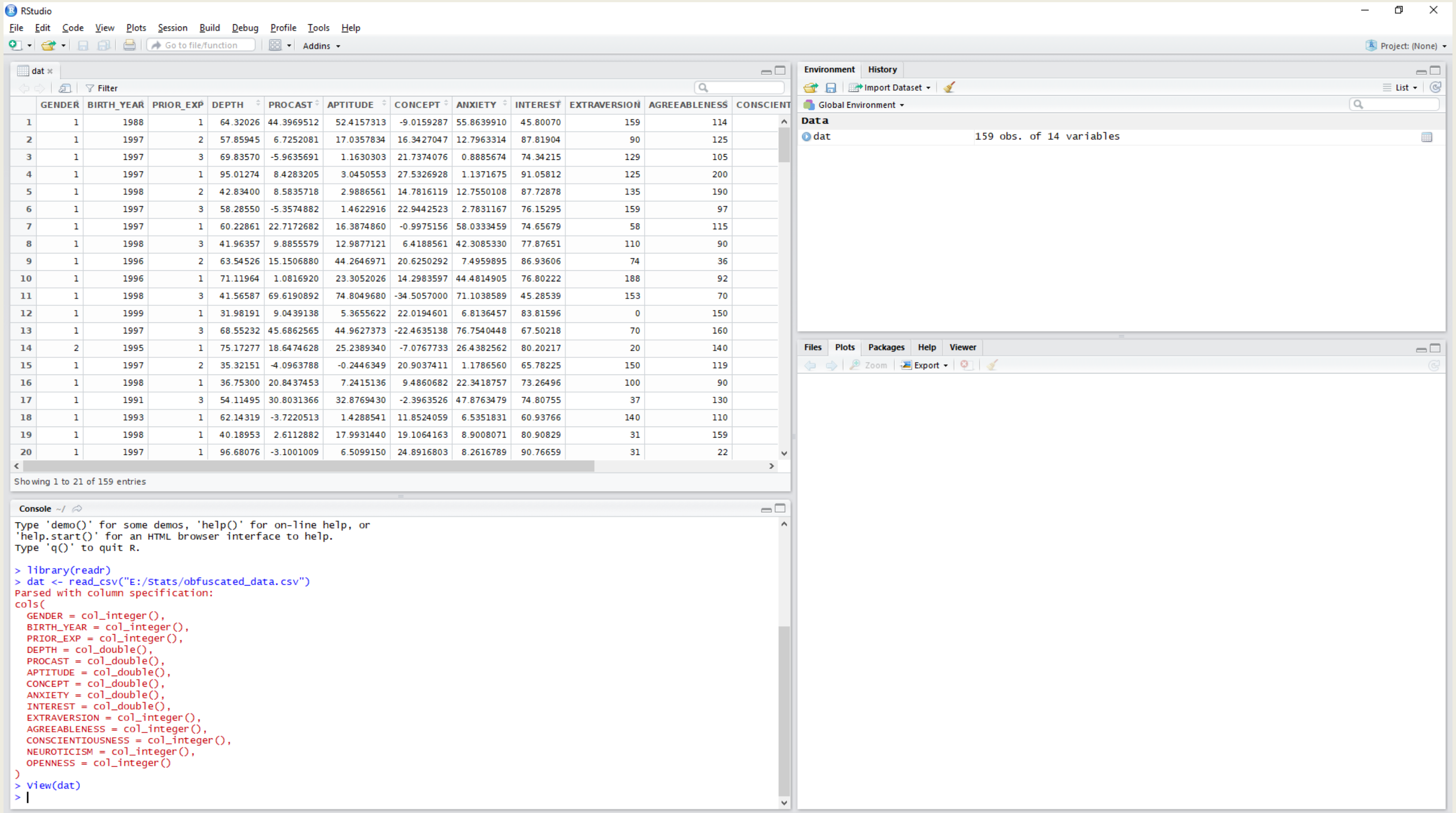
Import Data into R



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	8.4283205	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	58.0333459	74.65679	58	115	97	153	177
1	1998	3	41.96357	9.8855579	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	83.81596	0	150	110	5	40
1	1997	3	68.55232	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	5.0321124	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

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

Import Data into R



	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	2.9886561	14.7816119	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	58.0333459	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	83.81596	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	

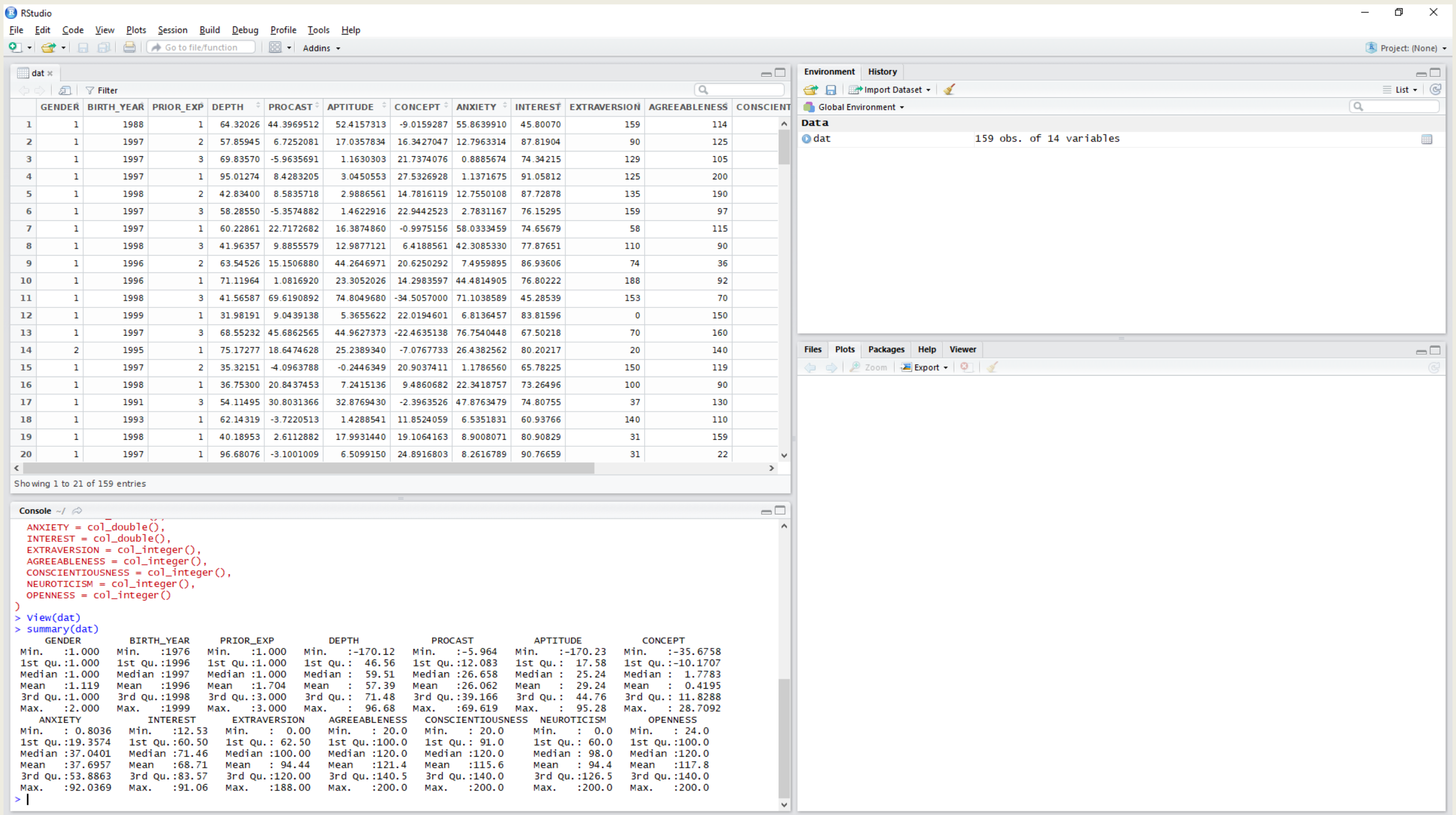
Showing 1 to 21 of 159 entries

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(readr)
> 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)
> 
```

GAMES ACADEMY FALMOUTH UNIVERSITY

Import Data into R



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

To see a correlation:

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

Try:

```
> cor(dat$PROCAST, dat$APTITUDE, method=“pearson”)
```

What does “2.2e-16” mean?

What can we conclude from this data?

Correlation

	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	2.9886561	14.7816119	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	58.0333459	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	83.81596	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	

Showing 1 to 21 of 159 entries

Data
dat 159 obs. of 14 variables

```
package ‘mnormt’ successfully unpacked and MD5 sums checked
package ‘psych’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Adrir\AppData\Local\Temp\RtmpkfoY8i\downloaded_packages
> library(psych)
Warning message:
package ‘psych’ was built under R version 3.4.3
> 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$APTITUDE)

        Pearson's product-moment correlation

data:  dat$PROCAST and dat$APTITUDE
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

> 
```

T-Test

To see a correlation:

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

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

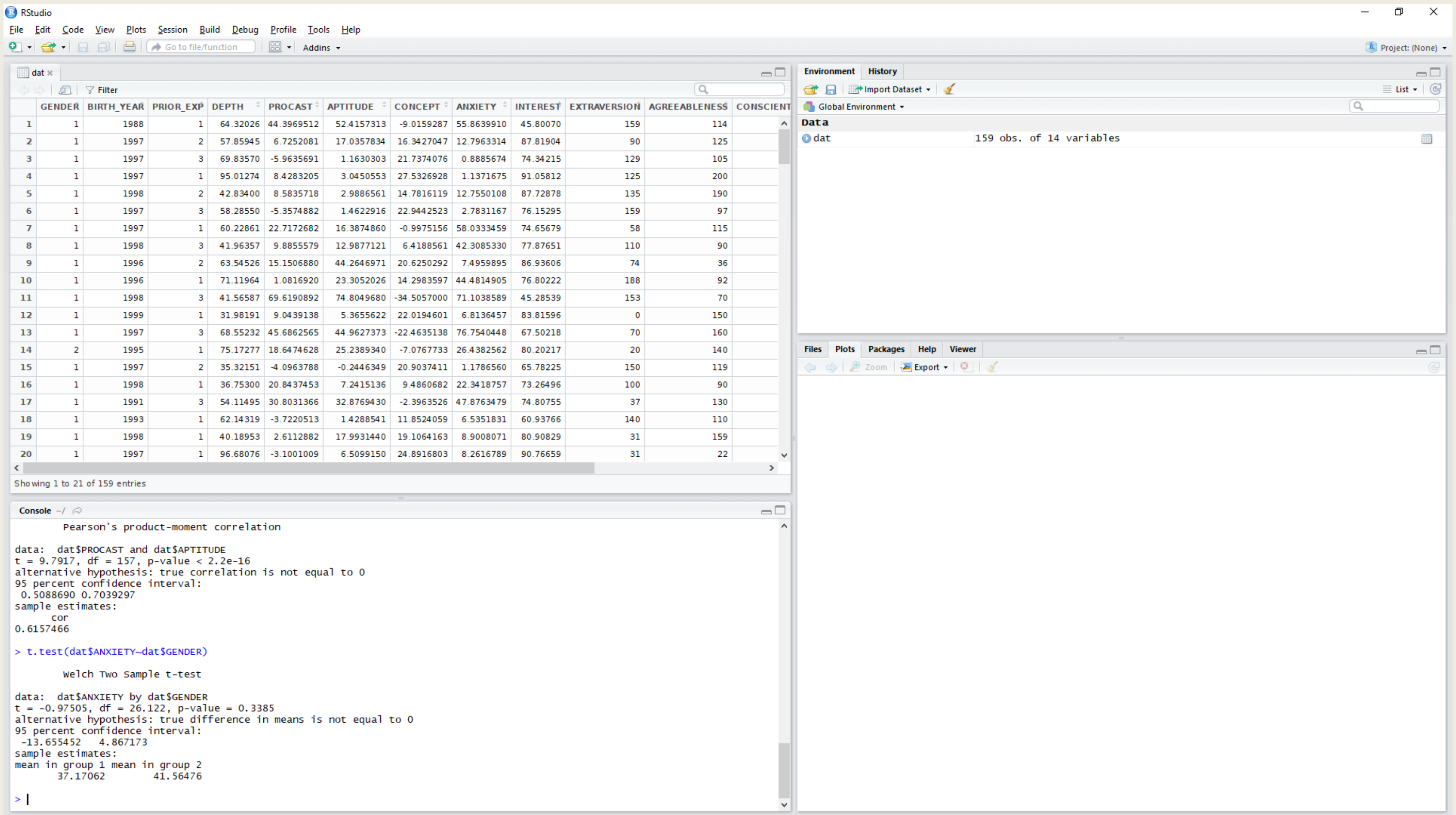
Try:

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

What can we conclude from this data?

GAMES ACADEMY | FALMOUTH UNIVERSITY

T-Test



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)
```

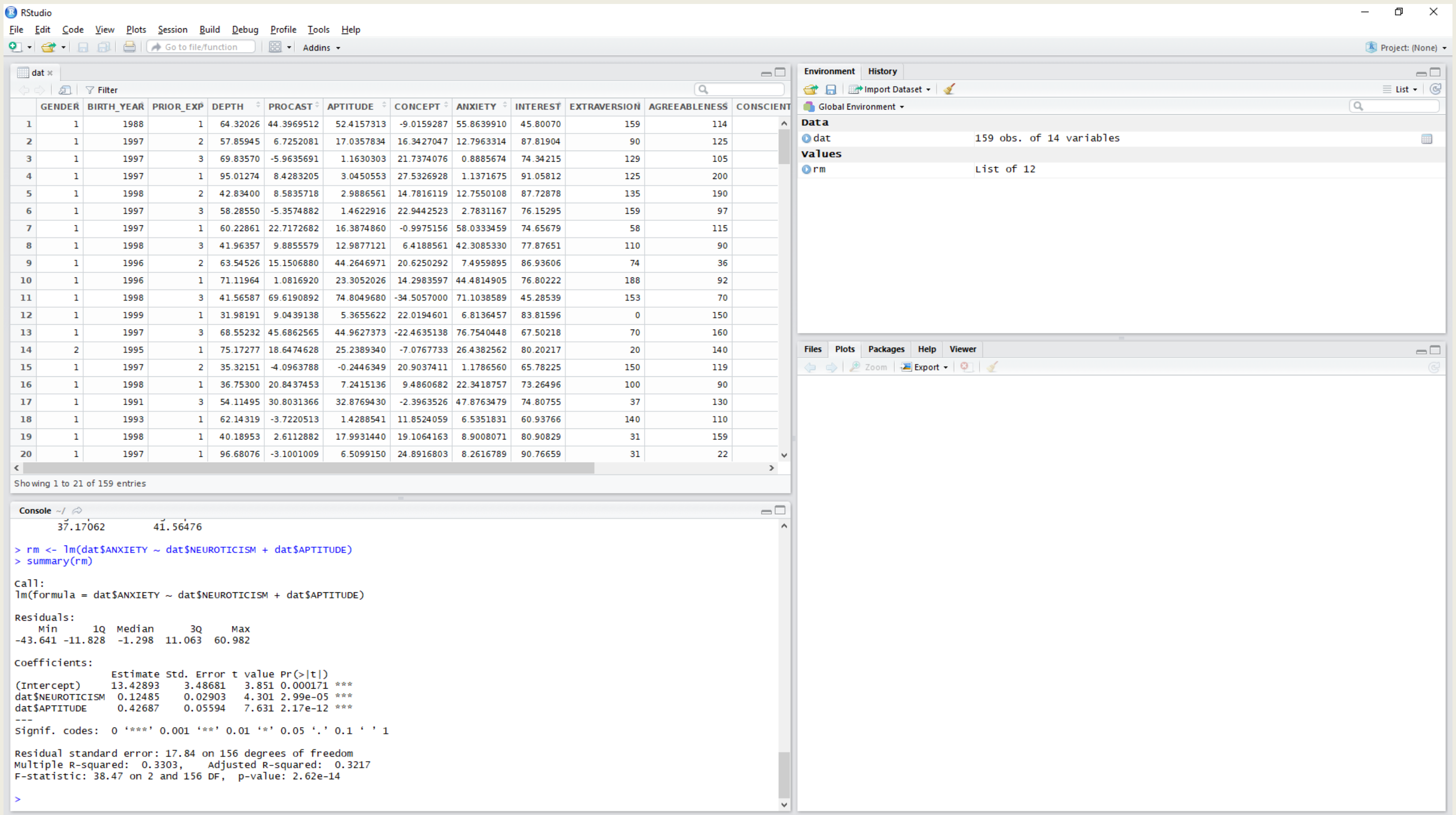
What does these results suggest?

To investigate relative importance of predictors, use:

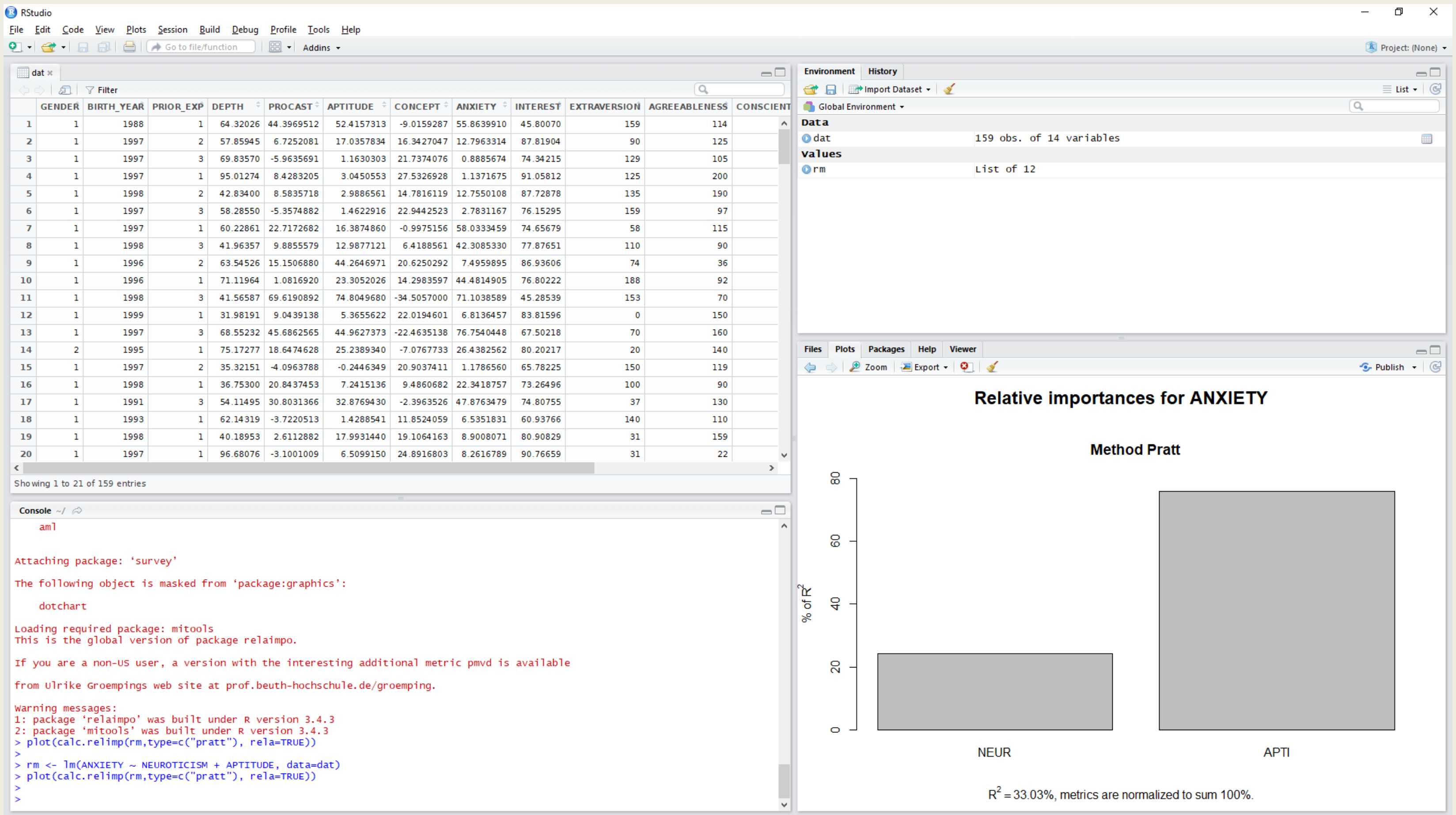
```
> library(relaimpo)
> calc.relimp(rm,type=c("pratt"), rela=TRUE)
```

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

Multiple Regression



Multiple Regression



Further Exploration in R

A Realm of Possibilities

Exploration

- 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.
  - Research the non-parametric statistical tests such as Mann-Whitney U and Kruskal-Wallis
  - Look for more advanced statistics which can handle different experimental designs such as ANCOVA

Reading

- Read and refer to the following:
  - www.rdocumentation.org
  - https://r4ds.had.co.nz/
  - www.statmethods.net
  - https://support.rstudio.com/hc/en-us
  - http://onlinestatbook.com/2/
  - https://web.stanford.edu/~hastie/ElemStatLearn/

Tutorials

Use Swirl for tutorials:

```
> install.packages(swirl)
> library(swirl)
> install_from_swirl(“R Programming”)
> swirl()
```