Statistical Analysis Presentation

A STUDY OF ALCOHOL CONSUMPTION IN ENGLAND

Introduction

Background

Alcohol is a significant public health issue, cited as “the leading cause of ill-health, disability, and death” for people in England aged 15-49 (Department of Health and Social Care, 2021).

Alcohol related hospital admissions in England have been rising year on year. 62% of admissions in 2018/19 were men, and admissions rise with age until 55-64 before reducing (NHS Digital, 2020).

Drinking amongst young women in the UK has been increasing, whilst conditions such as liver disease have also been increasing for the same group (Plant, 2008).

Secondary problems such as public order offences, accidents and injuries caused by binge drinking are more likely to be felt by men (Miller et al., 2005), whilst “higher percentages of heavier-typical quantity drinking are found in the younger age groups” (Chaiyasong et al., 2018).

Purpose of the presentation

This presentation uses data from Health Survey for England, 2011* to explore relationships and correlations between attributes such as alcohol consumption, gender, region, height and weight.

T-test, Chi-Square and Kruskal-Wallis tests for inferential statistics

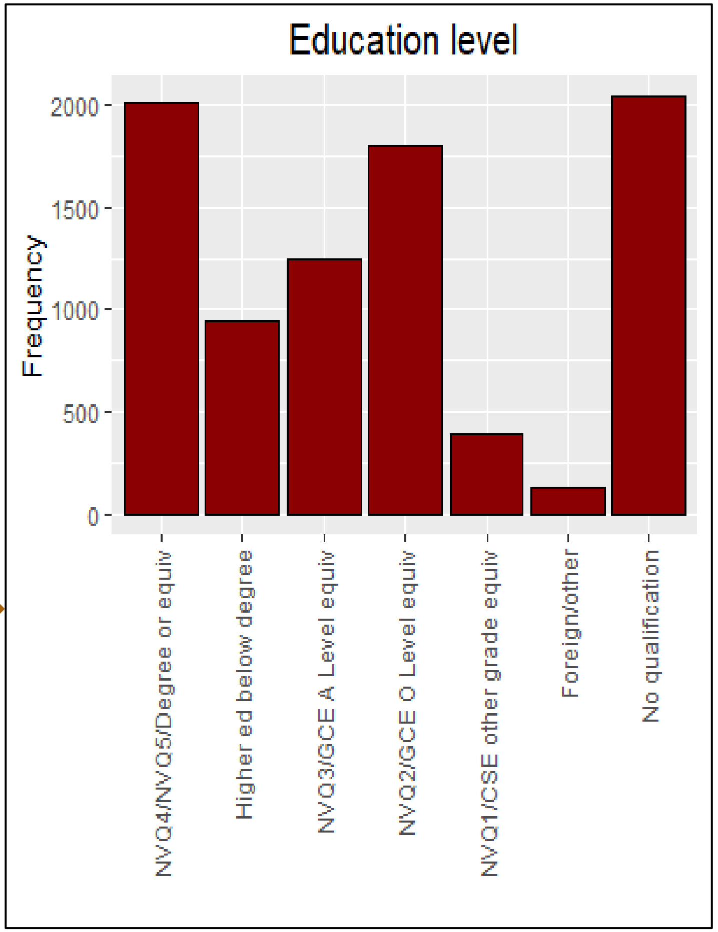
Shapiro-Wilk and Kolmogorov-Smirnov tests for distribution normality

R was used to perform the analysis because it is able to quickly perform statistical analysis and produce data visualisations.

*https://www.my-course.co.uk/pluginfile.php/891744/mod_assign/intro/HSE%202011.sav

Descriptive statistics

Question	Answer	Observation
How many people are included in the sample?	10,617	Very large sample
What is the percentage of people who drink alcohol?	63.2%	Figure includes N/A at 19.6% which is predominantly children under 18. If N/A is excluded it increases to 78.7%
What is the percentage of women in the sample?	54.3%	Similar to actual 50.9% females in the UK in 2011 (Clark, 2023)
What is the highest educational level?	No qualification (2,037)	‘No qualification’ had most responses. Highest ‘actual’ education level was NVQ4/NVQ5/Degree or equivalent:
What is the percentage of divorced people?	5.6%	Figure includes N/A at 18.9% If N/A is excluded it increases to 6.9%
What is the percentage of separated people?	2.1%	Figure includes N/A at 18.9% If N/A is excluded it increases to 2.6%



Descriptive statistics

	Household size	BMI	Age at last birthday
Mean	2.9	25.9	41.6
Median	3	25.6	42
Mode	2	Multi modal * See table	42 & 64
Minimum	1	8.34011	0
Maximum	10	65.27721	100
Range	1 - 10	8.34011 - 65.27721	0 - 100
Standard deviation	1.37	6.14	23.83

180 Modes for BMI								
13.76706	20.52755	22.44604	23.768	24.63076	25.3505	26.75475	27.91995	30.10691
15.07419	20.64521	22.47292	23.82012	24.66207	25.41852	26.79123	27.93738	30.15232
15.10565	20.96107	22.53362	23.82994	24.6755	25.44888	26.89171	28.03387	30.34498
15.11716	21.10538	22.76147	23.94858	24.72154	25.69331	26.97778	28.06514	30.55718
15.39402	21.16579	22.7878	23.97406	24.73146	25.77486	27.00408	28.11204	30.69307
15.58987	21.36725	22.82347	24.04789	24.7586	25.99036	27.02979	28.11526	30.81796
16.54064	21.54066	22.89376	24.0497	24.79339	26.03177	27.10957	28.12389	30.98742
16.56805	21.61622	22.90585	24.05754	24.80159	26.09568	27.11678	28.21979	31.04269
17.2607	21.62143	23.03436	24.11305	24.8826	26.10882	27.13884	28.27788	31.6068
17.62971	21.68367	23.08026	24.12092	24.96797	26.18647	27.19212	28.42288	31.93521
18.15462	21.72132	23.08843	24.27915	24.97874	26.23748	27.32885	28.46231	32.00386
18.38062	21.7333	23.18113	24.28338	25.02917	26.26407	27.37818	28.62897	32.2211
18.40747	21.74837	23.2767	24.29688	25.06672	26.2674	27.41375	28.65216	32.23528
19.2635	21.80385	23.28977	24.31359	25.10867	26.2701	27.42296	28.91054	32.28888
19.56781	22.04789	23.36088	24.35262	25.18414	26.34628	27.44727	28.9566	32.68683
19.7607	22.12727	23.39094	24.3598	25.22811	26.43323	27.49056	29.22573	32.88863
19.85798	22.13931	23.51504	24.46706	25.30832	26.50954	27.76621	29.88813	33.08129
20.01842	22.18928	23.56401	24.49239	25.32541	26.55283	27.82247	29.89543	33.3317
20.21217	22.21588	23.5782	24.5957	25.34064	26.65846	27.85467	29.93344	34.44429
20.43035	22.36915	23.73614	24.63003	25.34732	26.7128	27.91761	30.0971	36.84641

*BMI has 180 modes with just two instances each so could be considered no mode rather than multi-modal

Inferential statistics – which gender drinks more alcohol

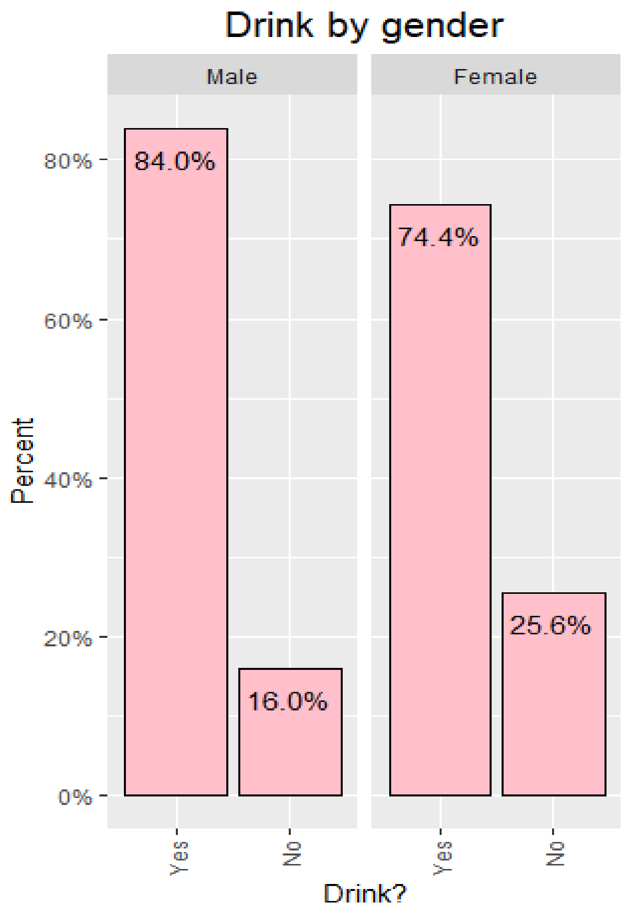
Chi-square test, because both variables are categorical, to see if there is a difference between whether males and females drink alcohol, α=0.05.

H_0: There is no difference between the percentage of men saying that they drink alcohol and women saying that they drink alcohol.

H_a: There is a difference between the percentage of men saying that they drink alcohol and women saying that they drink alcohol.

Gender	Does drink	Does not drink	Test value and p value
Male	3,172 (84.0%)	605 (16.0%)	Chi-square = 114.15 p-value < 2.2e-16
Female	3,540 (74.4%)	1,217 (25.6%)	

p-value < 0.05
We reject the null hypothesis
There is a difference between the percentage of men saying that they drink alcohol and women saying that they drink alcohol
A higher percentage of men drink than women



Inferential statistics – which gender drinks more alcohol

Test to see if there is a difference between the weekly quantity of alcohol consumed between males and females at 0.05 significance level (α=0.05).

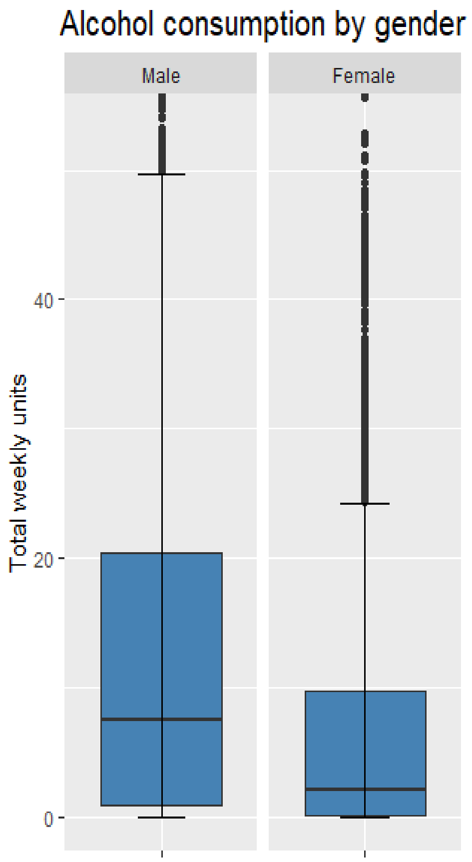
H_0: There is no difference in the average weekly quantity of alcohol consumed between men and women.

H_a: There is a difference in the average weekly quantity of alcohol consumed between men and women.

A Kolmogorov-Smirnov test was run plus a histogram and Q-Q plot examined of the total weekly units data and the distribution was not normal. However, due to the large sample size a t-test was still appropriate.

```
        Welch Two Sample t-test

data:  totalwu by as_factor(Sex)
t = 16.602, df = 5885.1, p-value < 2.2e-16
alternative hypothesis: true difference in means
 between group Male and group Female is not equal
 to 0
95 percent confidence interval:
 6.631599 8.407449
sample estimates:
  mean in group Male mean in group Female
           15.090413             7.570889
```

p-value < 0.05
We reject the null hypothesis
There is a difference in the average weekly quantity of alcohol consumed between men and women
The mean for men is 15.1 units
The mean for women is 7.6 units
The difference between the means at 95% confidence interval is 6.6 – 8.4 units



Inferential statistics – which region drinks the most alcohol

Chi-square test to see if there is a difference in the percentage of people who consume alcohol by region, α=0.05.

H_0 : There is no difference in the percentage of people who consume alcohol by region.

H_a : There is a difference in the percentage of people who consume alcohol by region.

Region	Does drink	Does not drink	Test value and p value
North East	576 (81.0%)	135 (19.0%)	Chi-square = 98.53 p-value < 2.2e-16
North West	833 (75.5%)	270 (24.5%)	
Yorkshire and The Humber	686 (77.3%)	201 (22.7%)	
East Midlands	624 (82.1%)	136 (17.9%)	
West Midlands	686 (76.8%)	207 (23.2%)	
East of England	763 (81.6%)	172 (18.4%)	
London	674 (68.9%)	304 (31.1%)	
South East	1,130 (81.6%)	255 (18.4%)	
South West	740 (83.9%)	142 (16.1%)	

p-value < 0.05
We reject the null hypothesis

There is a difference in the percentage of people who consume alcohol by region

South West consumes most alcohol

Inferential statistics – which region drinks the most alcohol

Test to see if there is a difference between the average weekly units of alcohol consumed between regions at 0.05 significance level ($\alpha=0.05$).

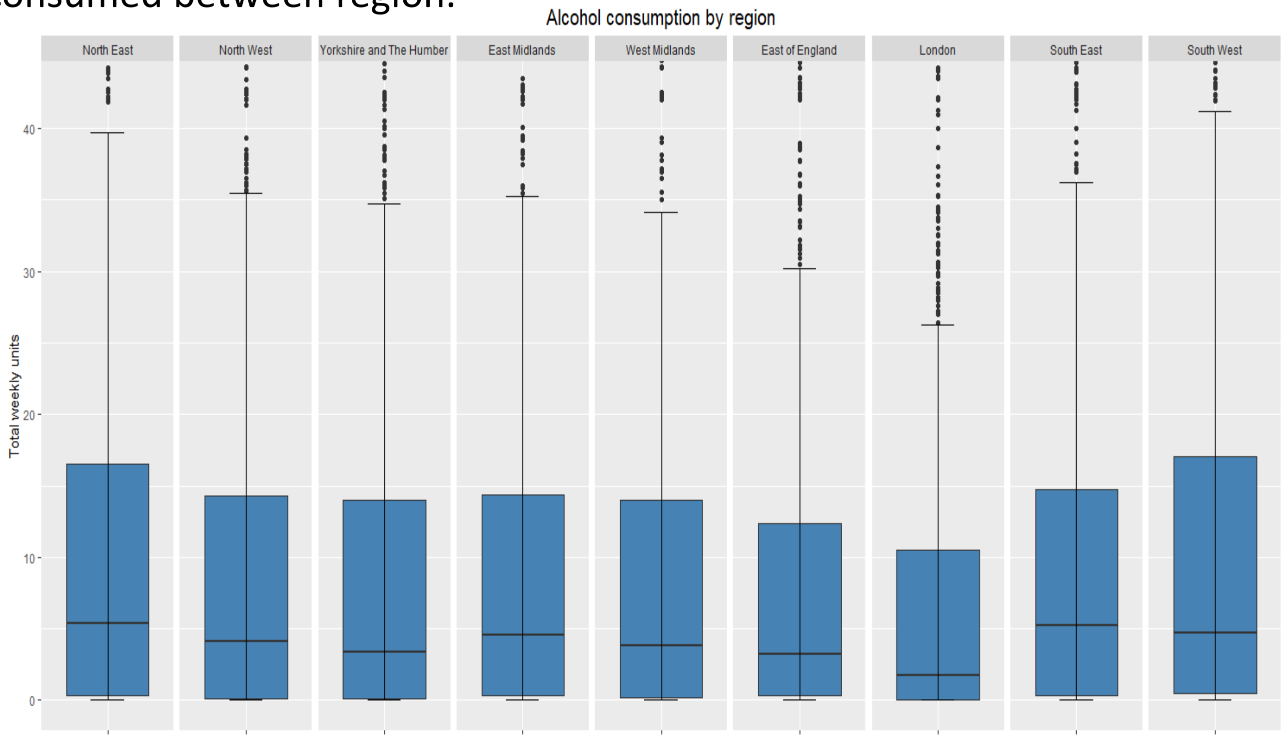
H_0: There is no difference in the average weekly quantity of alcohol consumed between regions.

H_a: There is a difference in the average weekly quantity of alcohol consumed between region.

Kruskal-Wallis test to compare multiple regions (groups) at the same time

```
        Kruskal-Wallis rank sum test

data:  totalwu by gor1
Kruskal-Wallis chi-squared = 81.722, df = 8, p-value = 2.199e-14
```

p-value < 0.05
We reject the null hypothesis
There is a difference in the average weekly quantity of alcohol consumed between regions



Is there is a statistical difference in height between men and women?

Subset of Adults created where Age > 17.

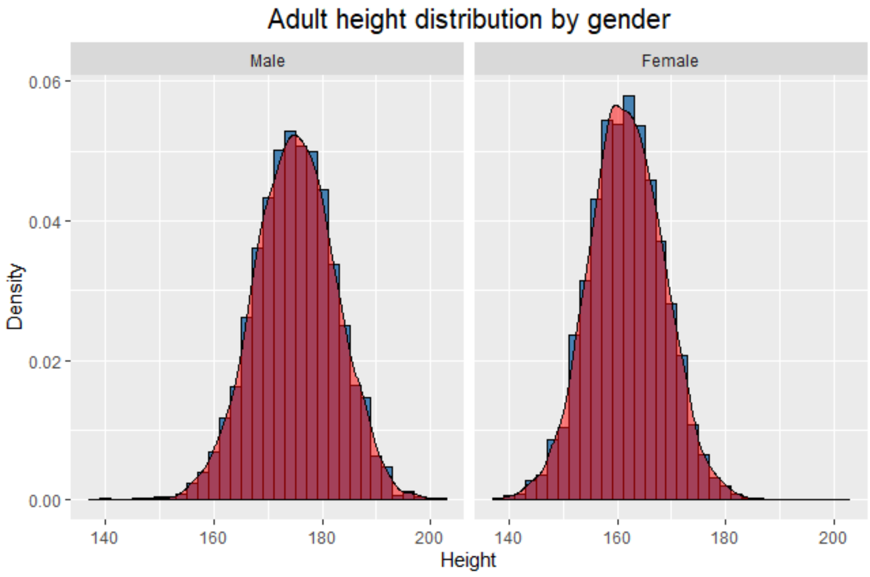
Shapiro-Wilk test to check for normal distributions in height of men and women.

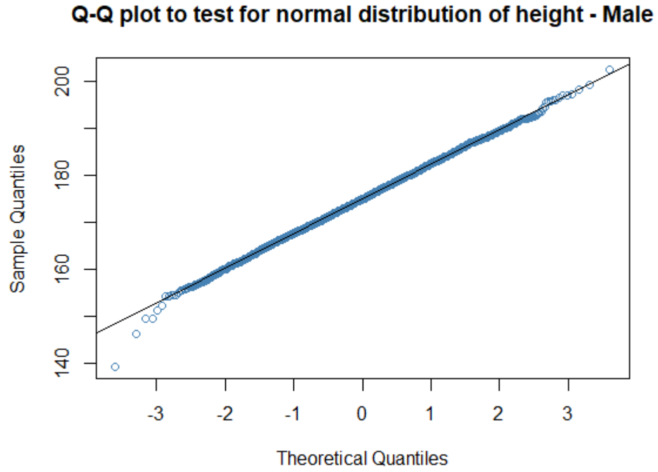
```
        Shapiro-Wilk normality test

data:  Adults$htval[as_factor(Adults$Sex) == "Male"]
W = 0.99919, p-value = 0.1674
```

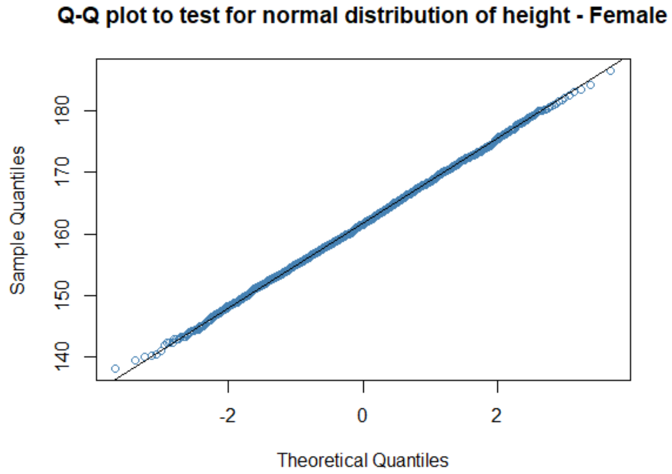
```
        Shapiro-Wilk normality test

data:  Adults$htval[as_factor(Adults$Sex) == "Female"]
W = 0.99961, p-value = 0.6573
```

With p-values > 0.05 both distributions are normal. The density plots and Q-Q plots also confirm normality.

A t-test was therefore be used.







Is there is a statistical difference in height between men and women?

H_0: There is no difference in height between men and women.

H_a: There is a difference in height between men and women.

```
        Welch Two Sample t-test

data:  htval by as_factor(Sex)
t = 77.707, df = 6444.4, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Male and group Female is not equal to 0
95 percent confidence interval:
 12.98930 13.66162
sample estimates:
  mean in group Male mean in group Female
            175.0300             161.7045
```

p-value < 0.05

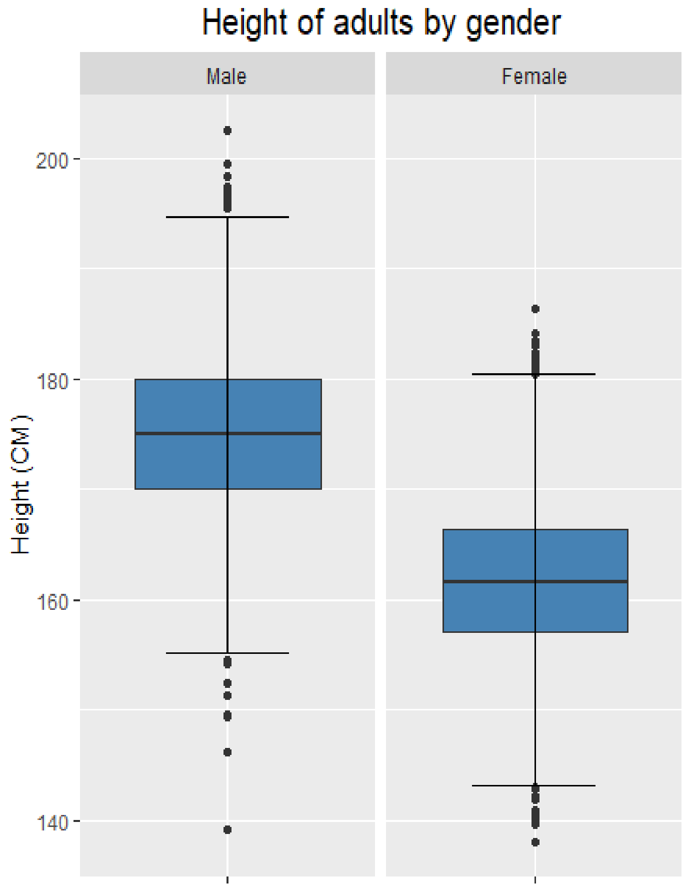
We reject the null hypothesis

There is a difference in height between men and women – men are taller

The mean for men in the sample is 175.0cm

The mean for women in the sample is 161.7cm

The difference between the means at 95% confidence interval is 13.0 – 13.7cm



Is there is a statistical difference in weight between men and women?

Shapiro-Wilk test to check for normal distributions in weight of men and women.

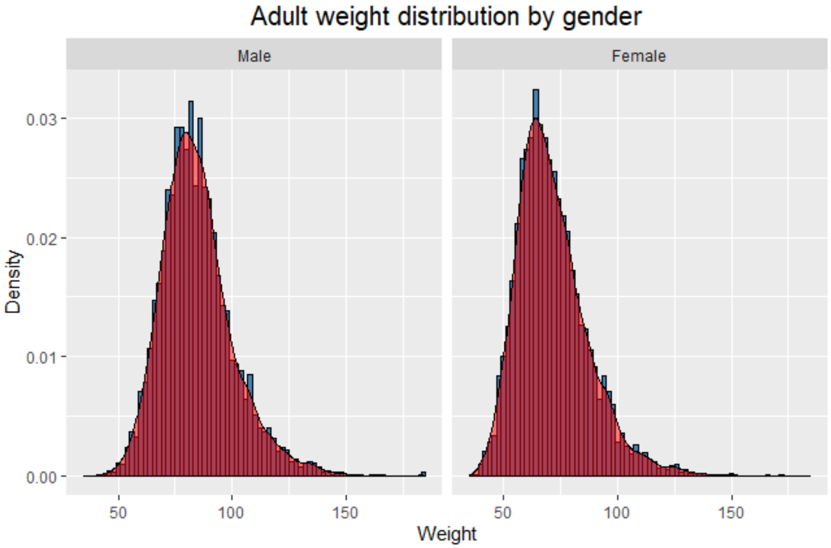
```
        Shapiro-Wilk normality test

data:  Adults$wtval[as_factor(Adults$Sex) == "Male"]
W = 0.95707, p-value < 2.2e-16
```

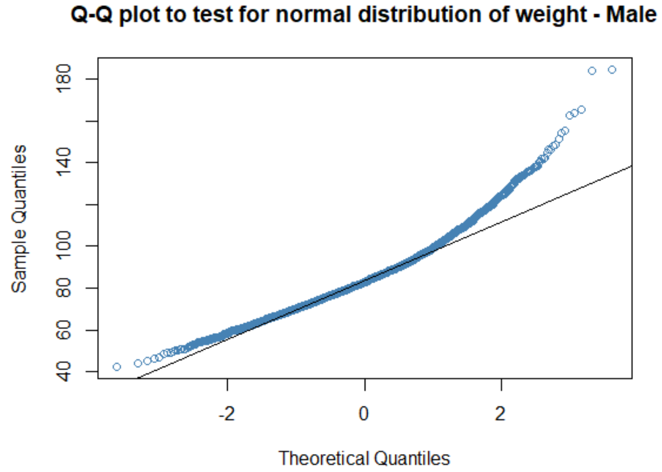
```
        Shapiro-Wilk normality test

data:  Adults$wtval[as_factor(Adults$Sex) == "Female"]
W = 0.94493, p-value < 2.2e-16
```

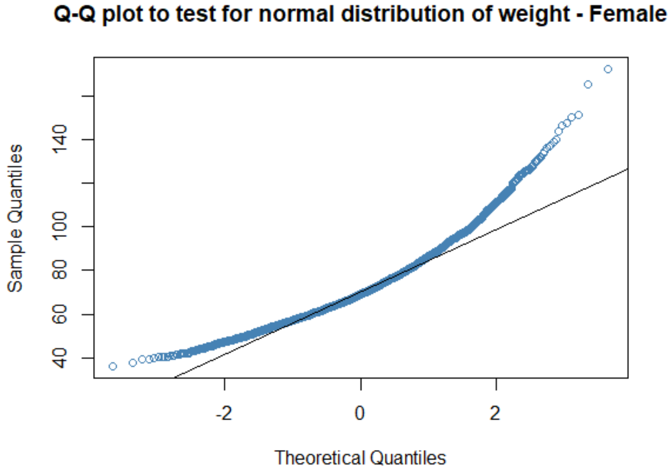
With p-values < 0.05, neither distribution was normal. The density plots were slightly skewed and the Q-Q plots deviated from normality.

With a large sample size the central limit theorem allows the use of a t-test even when the distribution is not normal.

A t-test was therefore used in favour of the non-parametric Mann-Whitney U test.







Is there is a statistical difference in weight between men and women?

H_0: There is no difference in weight between men and women.

H_a: There is a difference in weight between men and women.

```
        Welch Two Sample t-test

data:  wtval by as_factor(Sex)
t = 34.808, df = 6738.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Male and group Female is not equal to 0
95 percent confidence interval:
 12.57371 14.07447
sample estimates:
  mean in group Male mean in group Female
            84.89677             71.57267
```

p-value < 0.05

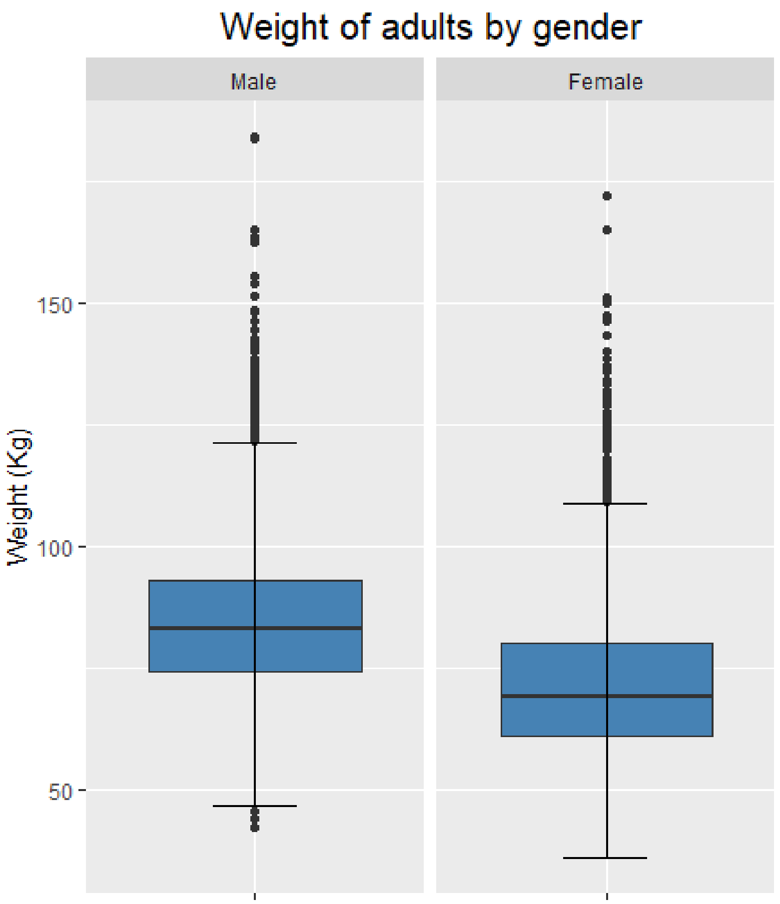
We reject the null hypothesis

There is a difference in height between men and women – men are heavier

The mean for men in the sample is 84.9kg

The mean for women in the sample is 71.6kg

The difference between the means at 95% confidence interval is 12.6 – 14.1kg



What is the correlation between whether a person drinks nowadays, total household income, age at last birthday and gender?

Kolmogorov-Smirnov tests to check for normal distributions for the four variables

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  HSE_2011$dnnow
D = 0.84134, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  HSE_2011$totinc
D = 0.99448, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  HSE_2011$Age
D = 0.95683, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  HSE_2011$Sex
D = 0.84134, p-value < 2.2e-16
alternative hypothesis: two-sided
```

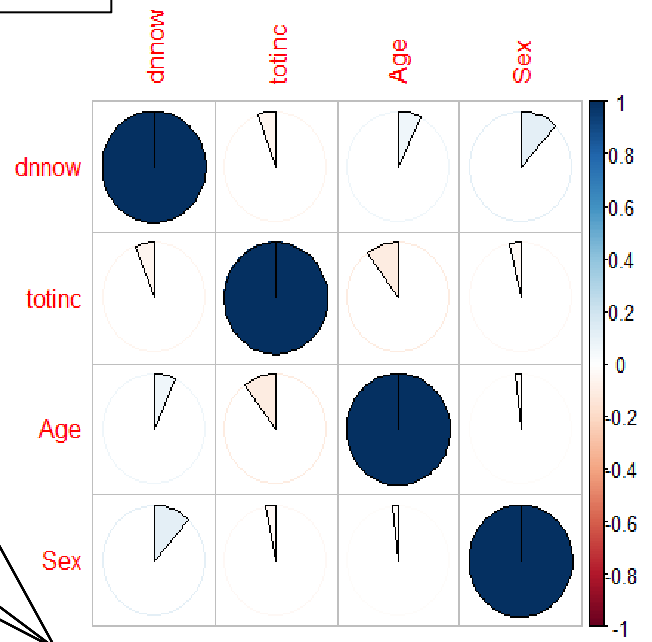
Histograms and Q-Q plots confirmed non-normal distributions so Spearman’s rank was used

What is the correlation between whether a person drinks nowadays, total household income, age at last birthday and gender?

p-value < 0.05
Small negative r value
Weak negative correlation

Strongest correlation

	Drink nowadays	Total household income	Age at last birthday	Gender
Drink nowadays		r_s= -0.06 p= 2.22e-07	r_s= 0.06 p= 2.506e-08	r_s= 0.12 p< 2.2e-16
Total household income	r_s= -0.06 p= 2.22e-07		r_s= -0.07 p= 5.324e-14	r_s= -0.03 p= 0.008748
Age at last birthday	r_s= 0.06 p= 2.506e-08	r_s= -0.07 p= 5.324e-14		r_s= 0.03 p= 0.001829
Gender	r_s= 0.12 p< 2.2e-16	r_s= -0.03 p= 0.008748	r_s= 0.03 p= 0.001829	



p-value < 0.05
Small positive r value
Weak positive correlation

Discussion and recommendations

The analysis found that men drink more than women, both when measured as number of people who drink, and average weekly units, consistent with the findings in NHS Digital (2020). Also consistent with the same study is that there is regional variation in levels of drinking. Men were also found on average to be taller and heavier than women.

Chaiyasong et al. (2018) found that high frequency drinking is higher in older age groups, and heavier drinking in middle and older age groups. This study found a weak positive correlation between drinking and age, but didn't differentiate heavier drinking.

Alcohol consumption amongst students is increasing (Davoren et al., 2016) and the study found that a high number of people are higher education educated which makes that demographic a useful target for raising awareness.

Recommendations

- Deliver a targeted campaign to increase awareness of the risks of alcohol throughout universities in England
- Deliver a wider public health campaign aimed at both genders but with specific elements targeted at men, focusing on the secondary effects of excessive alcohol consumption such as public disorder as well as primary health risks.

References

Chaiyasong, S. et al. (2018) Drinking patterns vary by gender, age and country-level income: Cross-country analysis of the International Alcohol Control Study. *Drug and Alcohol Review*, 37(S2): S53-S62. https://doi.org/10.1111/dar.12820

Clark, D. (2023) Population of the United Kingdom from 1953 to 2020, by gender. Available from: https://www.statista.com/statistics/281240/population-of-the-united-kingdom-uk-by-gender/ [Accessed 23 March 2023].

Davoren, M.P., Demant, J., Shiely, F. & Perry, I.J. (2016) Alcohol consumption among university students in Ireland and the United Kingdom from 2002 to 2014: a systematic review. *BMC public health*, 16(1): 1-13. https://doi.org/10.1186/s12889-016-2843-1

Department of Health and Social Care (2021) Delivering better oral health: an evidence-based toolkit for prevention. Chapter 12: Alcohol. Available from: https://www.gov.uk/government/publications/delivering-better-oral-health-an-evidence-based-toolkit-for-prevention/chapter-12-alcohol [Accessed 23 March 2023].

Miller, P., Plant, M. & Plant, M. (2005) Spreading out or concentrating weekly consumption: alcohol problems and other consequences within a UK population sample. *Alcohol and Alcoholism*, 40(5): 461-468. https://doi.org/10.1093/alcalc/agh169

NHS Digital (2020) Statistics on Alcohol, England 2020. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-alcohol/2020/part-1 [Accessed 23 March 2023].

Plant, M.L. (2008) The role of alcohol in women's lives: A review of issues and responses. *Journal of Substance Use*, 13(3): 155-191. https://doi.org/10.1080/14659890802040880

Appendix A: R code used to produce statistics and visualisations

```r
library(haven)
library(dplyr)
library(ggplot2)
library(corrplot)

HSE_2011 <- read_sav("HSE 2011.sav")
View(HSE_2011)

# How many people are included in the sample?
# pserial is the serial number of the individual, so count unique individuals
n_distinct(HSE_2011$pserial)

# What is the percentage of people who drink alcohol?
# Create a frequency table of whether people drink alcohol
table(
  as_factor(HSE_2011$dnnow)
)

# Make a proportional table that can be turned into percentages
prop.table(
  table(
    as_factor(HSE_2011$dnnow)
  )
)

# Repeat including the NA values
table(
  as_factor(HSE_2011$dnnow)
  ,useNA="ifany"
)

prop.table(
  table(
    as_factor(HSE_2011$dnnow)
    ,useNA="ifany"
  )
)

```

```r
# Plot percentage of people who drink alcohol
ggplot(HSE_2011, aes(as_factor(dnnow))) +
  geom_bar(aes(y=after_stat(count/sum(count))), colour='black' ,fill='pink') +
  labs(title='Percentage of people who drink alcohol', x='Drink alcohol?', y='Percent') +
  geom_text(aes( label = scales::percent(after_stat(count/sum(count))),
                 y=after_stat(count/sum(count))), stat= 'count', vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))


# Plot Male/female split
filter(HSE_2011, (dnnow > 0)) %>%
  ggplot(aes(as_factor(dnnow))) +
  geom_bar(aes(y=after_stat(prop), group=1), colour='black', fill='pink') +
  labs(title='Percentage people who drink alcohol, by gender', x='Drink alcohol?', y='Percent') +
  geom_text(aes(label = scales::percent(after_stat(prop)),
                y=after_stat(prop),group=1), stat= 'count', vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  facet_grid(~as_factor(Sex)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))


# What is the percentage of women in the sample?
# Create a frequency table of whether people drink alcohol
table(
  as_factor(HSE_2011$Sex)
)


# Make a proportional table that can be turned into percentages
prop.table(
  table(
    as_factor(HSE_2011$Sex)
  )
)*100

```

```r
# Plot all people
ggplot(HSE_2011, aes(as_factor(Sex))) +
  geom_bar(aes(y=after_stat(prop), group=1), colour='black', fill='pink') +
  labs(title='Gender breakdown', x='Gender', y='Percent') +
  geom_text(aes(label = scales::percent(after_stat(prop)),
                y=after_stat(prop),group=1), stat= 'count', vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

# Differentiate adults
age_labels <- c(
  'FALSE' = 'Under 18',
  'TRUE' = 'Adult'
)

# What is the highest educational level?
# Discover the attributes of topqual3
attributes(HSE_2011$topqual3)

# Now find the range
range(HSE_2011$topqual3, na.rm=TRUE)

table(as_factor(HSE_2011$topqual3))

filter(HSE_2011, topqual3 > 0) %>% # to remove NA values
  ggplot(aes(as_factor(topqual3))) +
  geom_bar(colour='black' ,fill='dark red') +
  labs(title='Qualifications', x='', y='Frequency') +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

# What is percentage of divorced and separated people?
# Plot all people
filter(HSE_2011, (marstatc > 0)) %>% # to remove NA values
  ggplot(aes(as_factor(marstatc))) +
  geom_bar(aes(y=after_stat(prop), group=1), colour='black', fill='pink') +
  labs(title='Marital status', x='', y='Percent') +
  geom_text(aes(label = scales::percent(after_stat(prop)),
                y=after_stat(prop),group=1), stat= 'count', vjust = -0.5) +
  scale_y_continuous(labels=scales::percent) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

```

```r
attributes(HSE_2011$marstatc)

table(
  as_factor(HSE_2011$marstatc)
  ,useNA="ifany"
)

prop.table(
  table(
    as_factor(HSE_2011$marstatc)
    ,useNA="ifany"
  )
)*100


# Find the mean, median, mode, minimum, maximum, range and standard deviation of household size,
# BMI and age at last birthday.
# First using the total sample
mean(HSE_2011$HHSize)
median(HSE_2011$HHSize)

# Modes function from https://stackoverflow.com/questions/2547402/how-to-find-the-statistical-mode
# Uses a function that calculated Mode, enhanced to remove NA values plus another that calculates
# multimodes. Combined here
Modes <- function(x, na.rm = FALSE) {
  if(na.rm){
    x = x[!is.na(x)]
  }
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}

Modes(HSE_2011$HHSize)
sort(table(HSE_2011$HHSize),decreasing=T)
min(HSE_2011$HHSize)
max(HSE_2011$HHSize)
range(HSE_2011$HHSize)
sd(HSE_2011$HHSize)
par(mar=c(5,5,4,4))
hist(HSE_2011$HHSize,main="Household size",xlab="Number of people",col="lightblue")

```

```r
ggplot(HSE_2011,
       aes(x=HHSize)) +
  geom_histogram(colour='black' ,fill='light blue', binwidth=1) +
  labs(title='Household size', x='Number of people', y='Frequency') +
  scale_x_continuous(breaks=1:10) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

SummaryDataHHS <- data.frame(Attribute = c('Mean', 'Median', 'Mode'),
                             Value = c(mean(HSE_2011$HHSize),
                                       median(HSE_2011$HHSize),
                                       Mode(HSE_2011$HHSize)
                             ))


ggplot(HSE_2011,
       aes(HHSize)) +
  geom_histogram(colour='black' ,fill='light blue', binwidth=1) +
  geom_vline(data=SummaryDataHHS,
             aes(xintercept=Value,
                 col=Attribute), size=1) +
  labs(title='Household size', x='Number of people', y='Frequency') +
  scale_x_continuous(breaks=1:10) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

SummaryDataHHS <- data.frame(Attribute = c('Mean', 'Mean -1 SD', 'Mean +1 SD'),
                          Value = c(mean(HSE_2011$HHSize),
                                    mean(HSE_2011$HHSize)-sd(HSE_2011$HHSize),
                                    mean(HSE_2011$HHSize)+sd(HSE_2011$HHSize)
                                    ))


ggplot(HSE_2011,
       aes(HHSize)) +
  geom_histogram(colour='black' ,fill='light blue', binwidth=1) +
  geom_vline(data=SummaryDataHHS,
             aes(xintercept=Value,
                 col=Attribute), size=1) +
  labs(title='Household size', x='Number of people', y='Frequency') +
  scale_x_continuous(breaks=1:10) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

```

```r
# Boxplot of household size
ggplot(HSE_2011,
       aes(HHSize)) +
  geom_boxplot(colour='black', fill='light blue') +
  stat_boxplot(geom ='errorbar', width=0.3) +
  coord_flip() +
  labs(title='Household size', x='Number of people') +
  scale_x_continuous(breaks=1:10) +
  scale_y_discrete() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))


# Need to remove NA values
mean(HSE_2011$bmival, na.rm=TRUE)
median(HSE_2011$bmival, na.rm=TRUE)
Modes(HSE_2011$bmival, na.rm=TRUE)
head(sort(table(HSE_2011$bmival),decreasing=T), n=200)
# We can see multiple BMI values with two counts

HSE_2011 %>% count(bmival) %>% arrange(desc(n))

min(HSE_2011$bmival, na.rm=TRUE)
max(HSE_2011$bmival, na.rm=TRUE)
range(HSE_2011$bmival, na.rm=TRUE)
sd(HSE_2011$bmival, na.rm=TRUE)
par(mar=c(5,5,4,4))
hist(HSE_2011$bmival,breaks=50,main="BMI value - 50 breaks",xlab="BMI",col="lightblue")

SummaryDataBMI <- data.frame(Attribute = c('Mean', 'Median'),
                             Value = c(mean(HSE_2011$bmival, na.rm=TRUE),
                                       median(HSE_2011$bmival, na.rm=TRUE)
                             ))


# Histogram of BMI
ggplot(HSE_2011,
       aes(bmival)) +
  geom_histogram(colour='black' ,fill='light blue', bins=50, na.rm = TRUE) +
  geom_vline(data=SummaryDataBMI,
             aes(xintercept=Value,
                 col=Attribute), size=1) +
  labs(title='BMI', x='BMI', y='Count') +
  scale_x_continuous(breaks = seq(from = 10, to = 100, by = 10)) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

```

```r
SummaryDataBMI <- data.frame(Attribute = c('Mean', 'Mean -1 SD', 'Mean +1 SD'),
                             Value = c(mean(rounded_bmi_1dp, na.rm=TRUE),
                                       mean(rounded_bmi_1dp, na.rm=TRUE)-sd(HSE_2011$bmival,
                                                                            na.rm=TRUE),
                                       mean(rounded_bmi_1dp, na.rm=TRUE)+sd(HSE_2011$bmival,
                                                                            na.rm=TRUE)
                             ))
# Histogram of BMI
ggplot(HSE_2011,
       aes(bmival)) +
  geom_histogram(colour='black' ,fill='light blue', bins=50, na.rm = TRUE) +
  geom_vline(data=SummaryDataBMI,
             aes(xintercept=Value,
                 col=Attribute), size=1) +
  labs(title='BMI - total sample', x='BMI', y='Frequency') +
  scale_x_continuous(breaks = seq(from = 10, to = 100, by = 10)) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

# Boxplot of BMI by gender
ggplot(HSE_2011,
       aes(bmival)) +
  geom_boxplot(colour='black', fill='light blue', na.rm = TRUE) +
  stat_boxplot(geom ='errorbar', width=0.3) +
  coord_flip() +
  labs(title='BMI value', x='') +
  scale_x_continuous(breaks=seq(0, 70, by=10)) +
  scale_y_discrete() +
  facet_grid(~as_factor(Sex)) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

mean(HSE_2011$Age)
median(HSE_2011$Age)
Modes(HSE_2011$Age)
# check with sorted table
sort(table(HSE_2011$Age),decreasing=T)
min(HSE_2011$Age)
max(HSE_2011$Age)
range(HSE_2011$Age)
sd(HSE_2011$Age)

par(mar=c(5,5,4,4))
hist(HSE_2011$Age,breaks=25,main="Age - 25 breaks",xlab="Age",col="lightblue")

```

```r
SummaryDataAge <- data.frame(Attribute = c('Mean', 'Median', 'Mode 1', 'Mode 2'),
                             Value = c(mean(HSE_2011$Age, na.rm=TRUE),
                                       median(HSE_2011$Age, na.rm=TRUE),
                                       Modes(HSE_2011$Age, na.rm=TRUE)
                             ))

ggplot(HSE_2011,
       aes(Age)) +
  geom_histogram(colour='black' ,fill='light blue', bins=101, na.rm = TRUE) +
  geom_vline(data=SummaryDataAge,
             aes(xintercept=Value,
                 col=Attribute), size=1.25) +
  labs(title='Age', x='Age', y='Frequency') +
  scale_x_continuous(breaks = seq(from = 0, to = 100, by = 10)) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

SummaryDataAge <- data.frame(Attribute = c('Mean', 'Mean -1 SD', 'Mean +1 SD'),
                             Value = c(mean(HSE_2011$Age, na.rm=TRUE),
                                       mean(HSE_2011$Age, na.rm=TRUE)-sd(HSE_2011$Age, na.rm=TRUE),
                                       mean(HSE_2011$Age, na.rm=TRUE)+sd(HSE_2011$Age, na.rm=TRUE)
                             ))

ggplot(HSE_2011,
       aes(Age)) +
  geom_histogram(colour='black' ,fill='light blue', bins=101, na.rm = TRUE) +
  geom_vline(data=SummaryDataAge,
             aes(xintercept=Value,
                 col=Attribute), size=1.25) +
  labs(title='Age', x='Age', y='Frequency') +
  scale_x_continuous(breaks = seq(from = 0, to = 100, by = 10)) +
  theme(plot.title = element_text(hjust = 0.5, size = 15))



# Inferential Statistics.
# Run a significance test to find out which gender drinks more alcohol.
# First using total weekly units as the measure of 'more alcohol'
# Check for normality using Kolmogorov-Smirnov test because the sample size is large
ks.test(HSE_2011$totalwu, 'pnorm')

hist(HSE_2011$totalwu,
     breaks=20,
     col='steelblue',
     main='Histogram to test for normal distribution of alcohol consumption',
     xlab='Units of alcohol per week')

```

```r
qqnorm(HSE_2011$totalwu,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of alcohol consumption')
qqline(HSE_2011$totalwu)

# Mann-Whitney U test to test for significance since the data is not normally distributed
wilcox.test(totalwu ~ Sex, data=HSE_2011)
t.test(totalwu ~ as_factor(Sex), data=HSE_2011)

# Next using ddnow as the measure of 'more alcohol'
# drink nowadays and gender are both binary values so use a contingency table and perform a
# chi-squared test.
table(as_factor(HSE_2011$dnnow), as_factor(HSE_2011$Sex))
chisq.test(table(HSE_2011$dnnow, HSE_2011$Sex))

filter(HSE_2011, (dnnow > 0)) %>% # to remove NA values
  ggplot(aes(as_factor(dnnow))) +
  geom_bar(aes(y=after_stat(prop), group=1), colour='black', fill='pink') +
  labs(title='Drink by gender', x='Drink?', y='Percent') +
  geom_text(aes(label = scales::percent(after_stat(prop)),
                y=after_stat(prop),group=1), stat= 'count', vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  facet_grid(~as_factor(Sex)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))


# Boxplot focusing on Q2, median and Q3
ggplot(HSE_2011, aes(x = '', y=totalwu)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.3) +
  labs(title = "Alcohol consumption by gender", y = 'Total weekly units', x='') +
  facet_grid(~as_factor(Sex)) +
  coord_cartesian(ylim = quantile(HSE_2011$totalwu, na.rm=TRUE, c(0.1, 0.97))) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

mean(HSE_2011$totalwu[as_factor(HSE_2011$Sex)=='Male'], na.rm=TRUE)
mean(HSE_2011$totalwu[as_factor(HSE_2011$Sex)=='Female'], na.rm=TRUE)
median(HSE_2011$totalwu[as_factor(HSE_2011$Sex)=='Male'], na.rm=TRUE)
median(HSE_2011$totalwu[as_factor(HSE_2011$Sex)=='Female'], na.rm=TRUE)

# Run a significance test to find out which region drinks the most alcohol.
# Use the Krushal-Wallis test to compare the alcohol consumption across the nine government regions:
kruskal.test(totalwu ~ gor1, data=HSE_2011)
```

```r

table(as_factor(HSE_2011$gor1), as_factor(HSE_2011$dnnow))

chisq.test(table(HSE_2011$gor1, HSE_2011$dnnow))

filter(HSE_2011, (dnnow > 0)) %>% # to remove NA values
  ggplot(aes(as_factor(dnnow))) +
  geom_bar(aes(y=after_stat(prop), group=1), colour='black', fill='pink') +
  labs(title='Drink by region', x='Drink?', y='Percent') +
  geom_text(aes(label = scales::percent(after_stat(prop)),
                y=after_stat(prop),group=1), stat= 'count', vjust = 2) +
  scale_y_continuous(labels=scales::percent) +
  facet_grid(~as_factor(gor1)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15)) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))


# Boxplot of alcohol by region focusing on Q1, median and Q3
ggplot(HSE_2011, aes(x = '', y=totalwu)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.3) +
  labs(title = "Alcohol consumption by region", y = 'Total weekly units', x='') +
  facet_grid(~as_factor(gor1)) +
  coord_cartesian(ylim = quantile(Adults$totalwu, na.rm=TRUE, c(0, 0.95))) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))


# Investigate whether there is a statistical difference between men and women on
# the following variables:
#   Valid height.
# Create a subset of adults
Adults <- subset(HSE_2011,Age > 17)
# Check to ensure only Adults are included:
range(Adults$Age)
# Now count the number of adults in the sample
n_distinct(Adults$pserial)

# Test for normal distribution
shapiro.test(Adults$htval[as_factor(Adults$Sex)=='Male'])
shapiro.test(Adults$htval[as_factor(Adults$Sex)=='Female'])
```

```r
ggplot(Adults,
       aes(x=htval)) +
  geom_histogram(aes(y=..density..), binwidth=2,
                 colour='black', fill='steel blue') +
  geom_density(alpha=.5, fill='red') +
  facet_grid(~as_factor(Sex)) +
  labs(title = "Adult height distribution by gender", y = 'Density', x='Height') +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

qqnorm(Adults$htval[as_factor(Adults$Sex)=='Male'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of height - Male')
qqline(Adults$htval[as_factor(Adults$Sex)=='Male'])

qqnorm(Adults$htval[as_factor(Adults$Sex)=='Female'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of height - Female')
qqline(Adults$htval[as_factor(Adults$Sex)=='Female'])

# t-test to test for significance
t.test(htval ~ as_factor(Sex), data=Adults)

ggplot(Adults, aes(x ='', y=htval)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.5) +
  labs(title = "Height of adults by gender", y = 'Height (CM)', x='') +
  facet_grid(~as_factor(Sex)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

#   Valid height.
# Test for normal distribution
shapiro.test(Adults$wtval[as_factor(Adults$Sex)=='Male'])
shapiro.test(Adults$wtval[as_factor(Adults$Sex)=='Female'])

ggplot(Adults,
       aes(x=wtval)) +
  geom_histogram(aes(y=..density..), binwidth=2,
                 colour='black', fill='steel blue') +
  geom_density(alpha=.5, fill='red') +
  facet_grid(~as_factor(Sex)) +
  labs(title = "Adult weight distribution by gender", y = 'Density', x='Weight') +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

```

```r
qqnorm(Adults$wtval[as_factor(Adults$Sex)=='Male'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of weight - Male')
qqline(Adults$wtval[as_factor(Adults$Sex)=='Male'])

qqnorm(Adults$wtval[as_factor(Adults$Sex)=='Female'],
       col='steelblue',
       main='Q-Q plot to test for normal distribution of weight - Female')
qqline(Adults$wtval[as_factor(Adults$Sex)=='Female'])

# Even though the distribution is not normal, we will still use the t-test
# because the sample size is large
t.test(wtval ~ as_factor(Sex), data=Adults)
#wilcox.test(wtval ~ Sex, data=Adults)

ggplot(Adults, aes(x = '', y=wtval)) +
  geom_boxplot(fill='steel blue') +
  stat_boxplot(geom ='errorbar', width=0.3) +
  labs(title = "Weight of adults by gender", y = 'Weight (Kg)', x='') +
  facet_grid(~as_factor(Sex)) +
  theme_grey() +
  theme(plot.title = element_text(hjust = 0.5, size = 15))

# What is the correlation between whether a person drinks nowadays, total
# household income, age at last birthday and gender?
# ks.test(HSE_2011$Sex, 'pnorm')

# drink nowadays is binary and total household income is ratio
# Kolmogorov-Smirnov because sample size is too large for Shapiro Wilk test
ks.test(HSE_2011$dnnow, 'pnorm')
ks.test(HSE_2011$totinc, 'pnorm')
ks.test(HSE_2011$Age, 'pnorm')
ks.test(HSE_2011$Sex, 'pnorm')

hist(HSE_2011$dnnow,
     breaks=100,
     col='steelblue',
     main='Histogram to test for normal distribution of drink nowadays',
     xlab='Units of alcohol per week')

hist(HSE_2011$totinc,
     breaks=100,
     col='steelblue',
     main='Histogram to test for normal distribution of total household income',
     xlab='Units of alcohol per week')
```

```r
hist(HSE_2011$Age,
     breaks=100,
     col='steelblue',
     main='Histogram to test for normal distribution of age',
     xlab='Units of alcohol per week')

hist(HSE_2011$Sex,
     breaks=100,
     col='steelblue',
     main='Histogram to test for normal distribution of gender',
     xlab='Units of alcohol per week')

qqnorm(HSE_2011$dnnow,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of drink nowadays')
qqline(HSE_2011$dnnow)

qqnorm(HSE_2011$totinc,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of total household income')
qqline(HSE_2011$totinc)

qqnorm(HSE_2011$Age,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of age')
qqline(HSE_2011$Age)

qqnorm(HSE_2011$Sex,
       col='steelblue',
       main='Q-Q plot to test for normal distribution of gender')
qqline(HSE_2011$Sex)

# Not normal distribution so use spearman's rank
cor.test(HSE_2011$dnnow, HSE_2011$totinc, method='spearman', use = 'complete.obs')
cor.test(HSE_2011$dnnow, HSE_2011$Age, method='spearman', use = 'complete.obs')
cor.test(HSE_2011$dnnow, HSE_2011$Sex, method='spearman', use = 'complete.obs')

cor.test(HSE_2011$totinc, HSE_2011$Age, method='spearman', use = 'complete.obs')
cor.test(HSE_2011$totinc, HSE_2011$Sex, method='spearman', use = 'complete.obs')

cor.test(HSE_2011$Age, HSE_2011$Sex, method='spearman', use = 'complete.obs')

cordata <- HSE_2011[,c('dnnow', 'totinc', 'Age', 'Sex')]
corrplot(cor(cordata, method='spearman', use = 'complete.obs'), method='pie')
```