```
---
title: "Student_Survey"
author: "Moshe Burnstein"
date: "`r Sys.Date()`"
output: pdf_document
---


```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
Calculate covariance of Student Survey

```{r    }
student_survey_df <- read.csv('http://content.bellevue.edu/cst/dsc/520/id/
resources/student-survey.csv', header = TRUE)
str(student_survey_df)
student_survey_matrix <- as.matrix(student_survey_df)
cov(student_survey_matrix)

```
**The covariance matrix shows the relationship between two variables. A
positive covariance indicates that as one variable goes, so does the other. A
negative covariance indicates that the variables go in opposite directions.
There appears to be a strong positive relationship between Happiness and
TimeTv. There appears to be an opposite relationship between TimeReading and
TimeTV and between TimeReading and Happiness. The covariance of Gender to all
other variables is minimal because Gender is represented by an integer of 1 or
0, so one cannot meaningfully compare.

The TimeReading variable seems to represent hours, while the TimeTv seems to
represent minutes. Happiness seems to be on a scale of 1 to 100...with its
range between 75 and 90 or so. Gender is arbitrarily assigned a 1 or a 0. One
can scale the variables to a standard deviation to get numbers between 0 and
1. If we standardize the covariance, we can assure a value between -1 and 1.**

```{r    }
library(Hmisc)
rcorr(student_survey_matrix)
cor(student_survey_matrix, method = "spearman")
rcorr(student_survey_matrix, type = "spearman")
cor(student_survey_matrix, method = "spearman")^2 * 100
```

Correlation test between 2 variables

```{r    }
cor.test(student_survey_df$TimeReading, student_survey_df$Happiness,
alternative = "less", method = "kendall")

```
```

I use Kendall's tau because because this is a small sample and I do not know if it is parametric. There are also tied ranks in TimeReading. I predict a negative correlation because the data shows that as TimeReading increases, so does Happiness decrease. The p-value of 0.12 tells us that we are only 88% probability that the results were not due to something random.
Correlation coefficient and coefficient of determination

```{r    }
cor.test(student_survey_df$TimeReading, student_survey_df$Happiness,
alternative = "less", method = "kendall", conf.level =  0.99)

(-0.2889428)^2 * 100

```

The -0.289 correlation coefficient indicates some negative correlation. Note that Kendall's coefficient is 66-75% less than Pearson's and Spearman, so consider the correlation to be solidly negative.

Watching more tv vs reading

```{r     }
tv_vs_read <- cor.test(student_survey_df$TimeTV,
student_survey_df$TimeReading, alternative = "less", method = "spearman")
tv_vs_read
(-0.9072536)^2 * 100


```
The p-value of 5.761e-05 indicates that the rho of -0.907 is meaningful and there is a very strong negative correlation. The r^2 of 82% indicates that 82% of the variables' variability is correlated to the other variable. However we cannot discern that watching more tv causes students to read less any more than reading less causes students to watch more tv.

Show correlation between TimeTv and TimeReading, controlling for Happiness

```{r      }
library(ggm)
par_cor <- pcor(c("TimeTV", "TimeReading", "Happiness"),
var(student_survey_df))
par_cor
par_cor^2
pcor.test(par_cor, 1, 11)

```
While the correlation between TimeTV and TimeReading is reduced when controlling for Happiness, there is still a strong negative correlation of 0.873. The R^2 is 76% which shows that 76% of the variance is due solely to the relationship between TimeTV and TimeReading. The p-value is less than 0.001, which indicates significance.