## Course: Visual Analytics for Policy and Management

# 

### Final Project: Broadband Internet Access in Stevenson, WA

# 

#### By Nathan Anderson-Hobbs, Johnson Nguyen, Kim Pearson & Marcela V?zquez

# 

# 

### Background

Stevenson, WA, is a small town of about 1500 people in the Portland Metropolitan Region. Although the small city features various natural and cultural amenities, many of its residents, businesses, and organizations have expressed dissatisfaction with local internet services. In February 2018, Stevenson City Council directed city staff to pursue the creation of a strategic plan regarding broadband infrastructure.

Broadband is high-speed internet, or internet with minimum speeds of 25 megabits per second (mbps) downloading and 3 mbps uploading, according to the US Federal Communications Commission (FCC). One of our team members, Kimberly Pearson, when interning at the City of Stevenson as the Broadband Project Coordinator, created a survey to assess current internet quality and uptake for busiesses and organizations. After sending the survey to over 150 small and medium sized businesses within city limits, 40 businesses responded, a ~26% response rate. The 40 survey questions refer to the business’ chosen Internet Service Provider, the cost of their internet service, the speeds they subscribe to, their actual experienced speeds, their satisfaction with their internet plan, and more. The results were then visualized in Tableau (see plots below).

###### Original plots: <https://public.tableau.com/profile/kimberly.pearson#!/vizhome/StevensonBusinessSurvey/Story1>

Within this report, we would like to explore the following research questions:

Additionally, we provide a tutorial for the following data visualizations:

1. Univariate

* Download Speed Range by ISP
* Upload Speed Range by ISP

1. Bivariate

* Subscribed Download Speed vs. Actual Speed
* Subscribed Upload Speed vs. Actual Speed

1. Multivariate

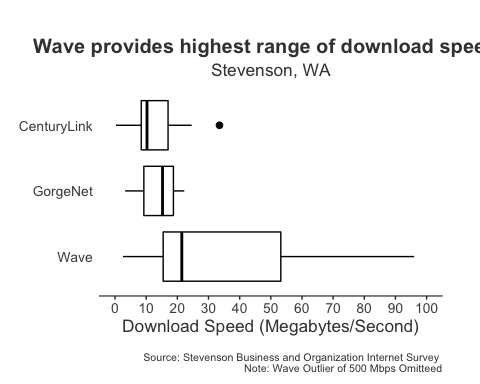
* Regression Plot for Speed Satisfaction
* Heat Map for Download Speed
* Map of Washington Fixed Broadband

### Univariate Plots

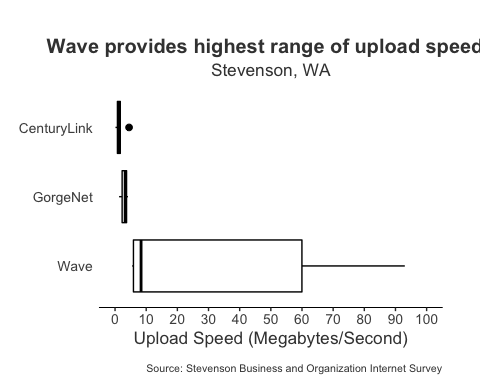
## Research Question:

## Explanation:

# Here we load the dataset.   
link='https://github.com/marcevl/Team-Assignments/raw/master/Internet.RData'  
load(file=url(link))  
  
# Then, we call the Ggplot2 library.   
library(ggplot2)  
  
# We create a box plot for each ISP, illustrating the ranges of their dowload speeds.   
box1 = ggplot(Internet, aes(x = ISP, y=Download.Speed))  
  
# We add information regarding the colors of the plot, and then flip it so the axes are reversed.   
box2 = box1 + geom\_boxplot(color='black',fill='white',  
 outlier.color='black',  
 outlier.size=2) + coord\_flip()  
   
# Here, we add a title, caption, and change the labels on each axis.   
box3 = box2 + labs(title='Wave provides highest range of download speeds', subtitle = 'Stevenson, WA',  
 y = 'Download Speed (Megabytes/Second)',  
 x = NULL,  
 caption = 'Source: Stevenson Business and Organization Internet Survey   
 Note: Wave Outlier of 500 Mbps Omitteed')  
  
# Then, we format the location, size, color, and type style of each text element.   
box4 = box3 + theme(plot.margin=unit(c(1,1,.5,.5),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0.5)),  
 plot.subtitle = (element\_text(hjust = 0.5)),  
 plot.caption=element\_text(size=8, vjust = -2),  
 text=element\_text(color='gray25', size = 13),  
 panel.background = element\_rect(fill='white'),  
 axis.ticks.y=element\_blank(),  
 axis.line.x=element\_line(color='black'))  
  
# Finally, we change the scale of each axis so that we show the exact number of tick marks that we'd like.   
box5 = box4 + scale\_y\_continuous(breaks=seq(0,100, 10),  
 limits=c(0,100)) + scale\_x\_discrete(limits=c('Wave','GorgeNet','CenturyLink'))  
  
# Then, we "print" the vizualization.   
box5



# We created a second box plot for each ISP, illustrating the ranges of their upload speeds.   
box1 = ggplot(Internet, aes(x= ISP, y = Upload.Speed))   
  
# We add information regarding the colors of the plot, and then flip it so the axes are reversed.   
box2 = box1 + geom\_boxplot(color='black',fill='white',  
 outlier.color='black',  
 outlier.size=2) + coord\_flip()  
  
# Here, we add a title, caption, and change the labels on each axis.  
box3 = box2 + labs(title='Wave provides highest range of upload speeds', subtitle = 'Stevenson, WA',  
 y = 'Upload Speed (Megabytes/Second)',  
 x = NULL,  
 caption = 'Source: Stevenson Business and Organization Internet Survey')  
  
# Then, we format the location, size, color, and type style of each text element.   
box4 = box3 + theme(plot.margin=unit(c(1,1,.5,.5),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0.5)),  
 plot.subtitle = (element\_text(hjust = 0.5)),  
 plot.caption=element\_text(size=8, vjust=-2),  
 text=element\_text(color='gray25', size = 13),  
 panel.background = element\_rect(fill='white'),  
 axis.ticks.y=element\_blank(),  
 axis.line.x=element\_line(color='black'))  
  
# Finally, we change the scale of each axis so that we show the exact number of tick marks that we'd like.   
box5 = box4 + scale\_y\_continuous(breaks=seq(0,100, 10),  
 limits=c(0,100)) + scale\_x\_discrete(limits=c('Wave','GorgeNet','CenturyLink'))  
  
# Then, we "print" the visualization.   
box5



### Bivariate plots

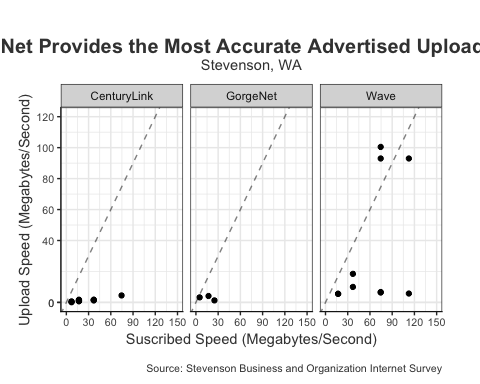
# For our second plots, we create a factor, ISP, with specific levels. It is based off the original variable.   
ISP <- factor(Internet$ISP)  
str(ISP)

## Factor w/ 6 levels "AT&T","CenturyLink",..: 2 3 6 2 3 2 1 6 6 3 ...

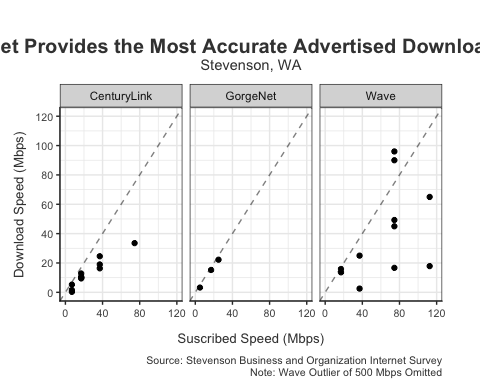
levels(ISP)

## [1] "AT&T" "CenturyLink" "GorgeNet" "Integra" "None"   
## [6] "Wave"

# Then, we create a dataframe with just the factors that we are concerned with in these plots. It contains information for all of the ISPs.   
df1 <- data.frame(ISP, Internet$Subscribed.Speed, Internet$Download.Speed, Internet$Upload.Speed)  
names(df1) <- c("ISP", "Subscribed", "Download", "Upload")  
  
# Next, we subsetted the data to only include Wave, CenturyLink, and GorgeNet.   
df2 <- subset(df1, ISP %in% c("Wave", "CenturyLink", "GorgeNet"))  
  
# We then created a base plot of Subscribed Speeds versus Upload Speeds.   
base = ggplot(df2, aes(x=Subscribed, y=Upload))  
  
# We added a scatterplot layer.   
box1 = base + geom\_point()  
  
# Then, a reference line that represents the subscribed speeds.   
box2 = box1 + geom\_abline(a=0,b=45,lty=2,lwd=0.5, alpha = 0.5)  
  
# We used the commadn facet wrap to separate the plot into three plots, one for each ISP.   
box3 = box2 + facet\_wrap( ~ ISP ,nrow = 1)  
  
# Then, we added a scale for the y axis and a scale for the x axis.   
box4 = box3 + scale\_y\_continuous(breaks=c(0,0,40,60,80,100,120),  
 limits = c(0, 120))  
  
box5 = box4 + scale\_x\_continuous(breaks=c(0,30,60,90,120,150),  
 limits = c(0, 150))  
  
# We coded for the title.   
Title = 'GorgeNet Provides the Most Accurate Advertised Upload Speeds'  
  
# We added the title, subtitle, x-axis label, y-axis label, and the caption.   
box6 = box5 + labs(title=Title, subtitle = 'Stevenson, WA',  
 x= 'Suscribed Speed (Megabytes/Second)',  
 y = 'Upload Speed (Megabytes/Second)',  
 caption='Source: Stevenson Business and Organization Internet Survey') + theme\_bw()  
  
# We formatted for each text element, as before.   
box7 = box6 + theme(plot.margin=unit(c(1,1,.5,.5),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0.5)),  
 plot.subtitle = (element\_text(hjust = 0.5, vjust = 2)),  
 plot.caption=element\_text(size=8, vjust=-2),  
 text=element\_text(color='gray25', size = 11),  
 axis.line.y=element\_line(color = 'black'),  
 axis.line.x=element\_line(color='black'),  
 axis.text=element\_text(size=8))  
  
# Finally, we "printed" the faceted scatterplot.  
box7



# The instructions to create the faceted scatterplot are the same as in the past plot.   
base = ggplot(df2, aes(x=Subscribed, y=Download))  
  
box1 = base + geom\_point()  
  
box2 = box1 + geom\_abline(a=0,b=45,lty=2,lwd=0.5, alpha = 0.5)  
  
box3 = box2 + facet\_wrap( ~ ISP ,nrow = 1)  
  
box4 = box3 + scale\_y\_continuous(breaks=seq(0,120,20),  
 limits = c(0,120))  
  
box5 = box4 + scale\_x\_continuous(breaks=seq(0,120,40),  
 limits = c(0,120))  
  
Title = 'GorgeNet Provides the Most Accurate Advertised Download Speeds'  
  
box6 = box5 + labs(title=Title, subtitle = 'Stevenson, WA',  
 x= 'Suscribed Speed (Mbps)',  
 y = 'Download Speed (Mbps)',  
 caption='Source: Stevenson Business and Organization Internet Survey  
 Note: Wave Outlier of 500 Mbps Omitted') + theme\_bw()  
  
box7 = box6 + theme(plot.margin=unit(c(1,1,.5,.5),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0.5)),  
 plot.subtitle = (element\_text(hjust = 0.5, vjust = 2)),  
 plot.caption=element\_text(size=8, vjust=-3),  
 text=element\_text(color='gray25', size = 11),  
 axis.line.y=element\_line(color = 'black'),  
 axis.line.x=element\_line(color='black'),  
 axis.text=element\_text(size=8),  
 axis.text.x = element\_text(angle = 0, vjust = 0.7),  
 axis.title.x = element\_text(vjust = -2, size = 10),  
 axis.title.y = element\_text(size = 10, vjust = 2.5))  
box7



### Multivariate Plots

Research Question: Explanation:

# Here, we called three libraries that would help us make dot and whisker plots as well as   
library(dotwhisker)  
library(broom)  
library(dplyr)  
  
# We created the first model of regression for Download and Upload Speeds on Satisfaction with Speed  
model1=lm(Satisfaction.Speed~Download.Speed+Upload.Speed,data=Internet2[,-1])  
summary(model1)

##   
## Call:  
## lm(formula = Satisfaction.Speed ~ Download.Speed + Upload.Speed,   
## data = Internet2[, -1])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4111 -0.4159 0.4289 0.5825 1.5775   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.395876 0.219081 15.501 2.49e-14 \*\*\*  
## Download.Speed 0.001154 0.008049 0.143 0.887   
## Upload.Speed 0.002786 0.007791 0.358 0.724   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.043 on 25 degrees of freedom  
## (12 observations deleted due to missingness)  
## Multiple R-squared: 0.1219, Adjusted R-squared: 0.05169   
## F-statistic: 1.736 on 2 and 25 DF, p-value: 0.1968

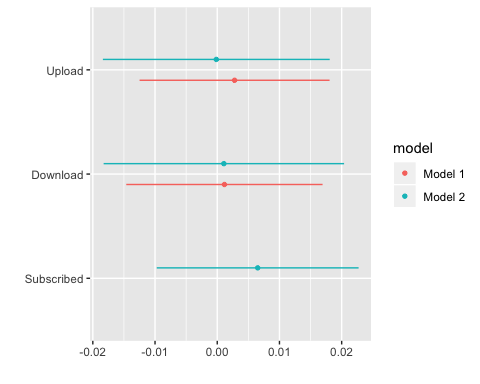
# Here we tidied up and mutated the model of regression so thatwe add a column, "model," with the values of Model 1.   
model1\_t = tidy(model1) %>% mutate(model = "Model 1") %>%  
 relabel\_predictors(Upload.Speed = "Upload",  
 Download.Speed = "Download",  
 Subscribed.Speed = "Subscribed")

# We then created a second model of regression, this time with Subscribed Speeds.   
model2=lm(Satisfaction.Speed~Download.Speed+Upload.Speed+Subscribed.Speed,data=Internet2[,-1]) # ~. means: all the remaining variables  
summary(model2)

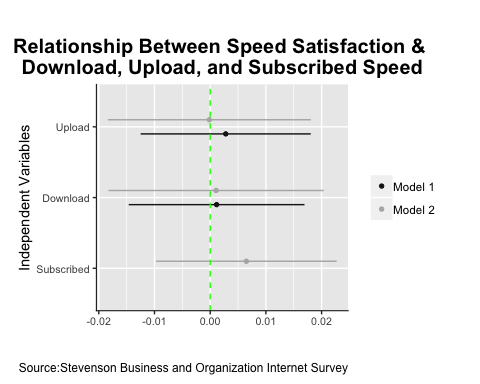
##   
## Call:  
## lm(formula = Satisfaction.Speed ~ Download.Speed + Upload.Speed +   
## Subscribed.Speed, data = Internet2[, -1])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3054 -0.3388 0.2492 0.6909 1.6888   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.1851440 0.3286583 9.691 1.38e-09 \*\*\*  
## Download.Speed 0.0010622 0.0098589 0.108 0.915   
## Upload.Speed -0.0001521 0.0093008 -0.016 0.987   
## Subscribed.Speed 0.0064991 0.0082764 0.785 0.440   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.069 on 23 degrees of freedom  
## (13 observations deleted due to missingness)  
## Multiple R-squared: 0.145, Adjusted R-squared: 0.03353   
## F-statistic: 1.301 on 3 and 23 DF, p-value: 0.2981

# We tidied up and mutated the second model of regression to have a column called "model," with values of "Model 2"  
model2\_t <- tidy(model2) %>% mutate(model = "Model 2") %>%  
 relabel\_predictors(Upload.Speed = "Upload",  
 Download.Speed = "Download",  
 Subscribed.Speed = "Subscribed")

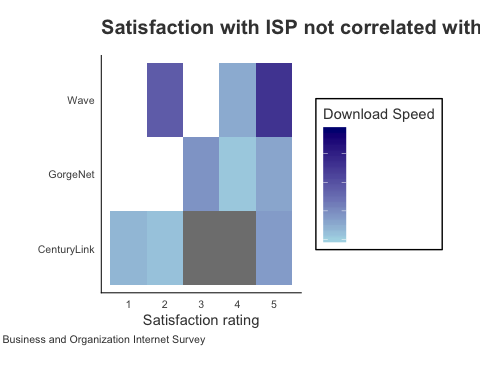
# Here we combined all the models   
allModels=rbind(model1\_t, model2\_t)  
  
# Then we plotted all the models.   
dwplot(allModels)



# Here we plot all the models and add or change theme/text elements   
dwplot(allModels) +   
 geom\_vline(xintercept = 0,   
 colour = "green",   
 linetype = 2) +   
 scale\_colour\_grey(start = .1, end = .7) +   
 labs(y = "Independent Variables",  
title = "Relationship Between Speed Satisfaction &   
Download, Upload, and Subscribed Speed",   
 caption = "Source:Stevenson Business and Organization Internet Survey") +  
 theme(plot.margin=unit(c(1,1,.5,.5),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0.5)),  
 plot.subtitle = (element\_text(hjust = 0.5, vjust = 2)),  
 plot.caption=element\_text(size=9, vjust=-2),  
 text=element\_text(color='black', size = 11),  
 legend.title = element\_blank(),  
 axis.line.y=element\_line(color = 'black'),  
 axis.line.x=element\_line(color='black'),  
 axis.text=element\_text(size=8))



##Heatmap  
  
library(scales)  
library(ggplot2)  
  
#loading data  
link='https://github.com/marcevl/Team-Assignments/raw/master/Internet2.RData'  
load(file=url(link))  
  
#changing download speed to a variable from 0-1  
Internet2$Download.Speed.Rescaled <- rescale(Internet2$Download.Speed)  
  
#adding data to data frame  
df1 <- Internet2  
  
#removing obervation with outlier  
df1 <- df1[-3,]  
  
#removing unnecessary variables from heatmap data frame  
df.heatmap <- df1[c(2,6,10,40)]  
  
#removing ISPs without many observations  
df.heatmap <- subset(df.heatmap, ISP %in% c("Wave", "CenturyLink", "GorgeNet"))  
  
#plotting heat map  
heat1 = ggplot(data = df.heatmap, aes(x = Satisfaction.Speed, y = ISP)) +  
 geom\_tile(aes(fill = Download.Speed.Rescaled))  
  
heat1 = heat1 + labs(title='Satisfaction with ISP not correlated with Download Speed',  
 y=NULL,  
 x='Satisfaction rating',  
 caption='Source: Stevenson Business and Organization Internet Survey',  
 fill='Range of \nDownload Speeds')  
  
heat2 = heat1 + theme(plot.margin=unit(c(.5,1,1,1),'cm'),  
 plot.title=(element\_text(face='bold',  
 size=15,  
 hjust=0,  
 margin=unit(c(0,0,.5,0),'cm'))),  
 plot.subtitle = (element\_text(hjust = 0.5, vjust = 2)),  
 plot.caption=element\_text(size=8, hjust=2),  
 legend.text=element\_blank(),  
 legend.background=element\_rect(linetype='solid',  
 color='black'),  
 panel.background=element\_rect(fill='white'),  
 text=element\_text(color='gray25', size = 11),  
 axis.line.y=element\_line(color = 'black'),  
 axis.line.x=element\_line(color='black'),  
 axis.ticks=element\_blank(),  
 axis.text=element\_text(size=8))  
  
heat2 + scale\_fill\_gradient(low='lightblue',high='navy',  
 guide\_colorbar(reverse=TRUE,  
 title='Download Speed'))



### Conclusion

To conclude….