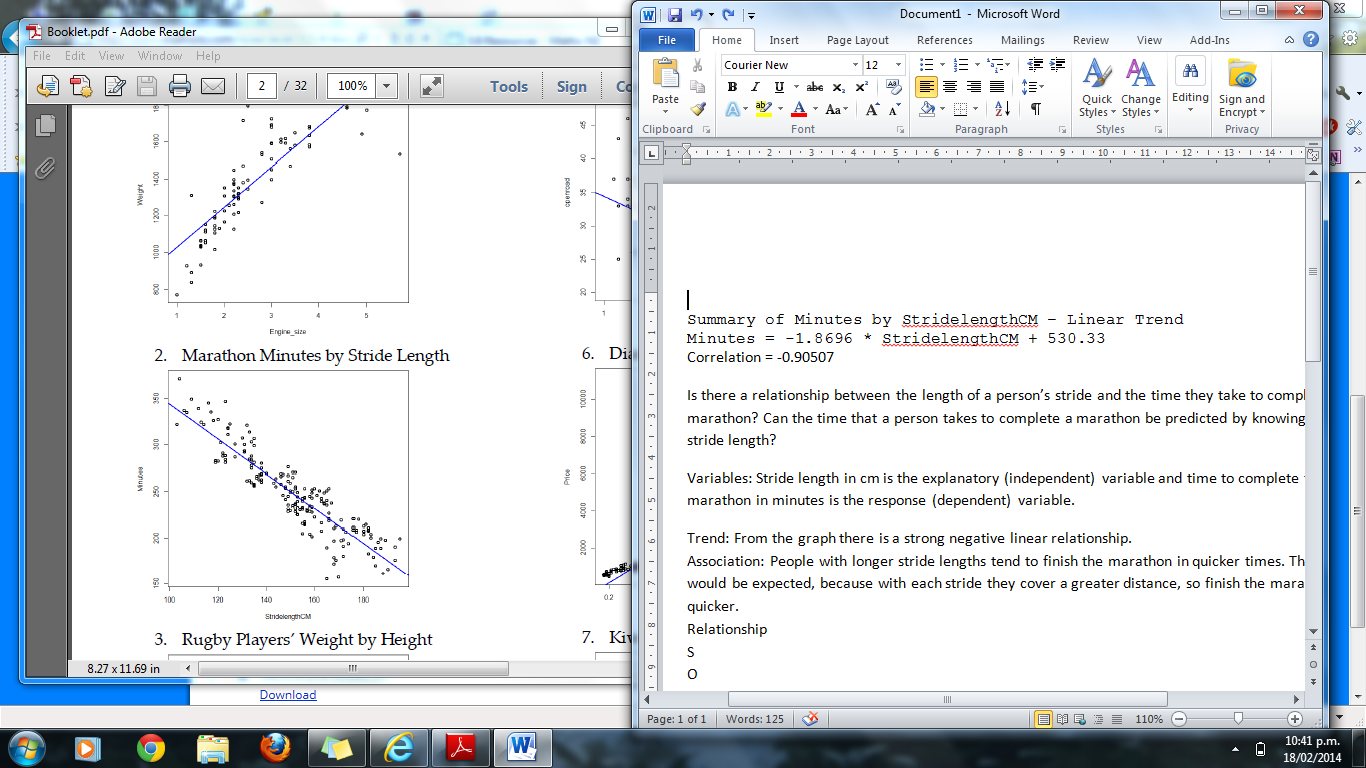
**2. Marathon**

The data is a sample taken from marathons in NZ.

|  |  |
| --- | --- |
| It is a simple random sample of 200 athletes. **Variable** | **Description** |
| Minutes | How many minutes they completed the marathon in |
| Gender | Male (M) or Female (F) |
| AgeGroup | Younger (under 40) or older (over 40) |
| StridelengthCM | The persons average stride length over the marathon in cm. |



Summary of Minutes by StridelengthCM

Linear Trend

Minutes = -1.8696 \* StridelengthCM + 530.33

Correlation = -0.90507

I wonder if there is a relationship between the length of a person’s stride and the time they take to complete a marathon. Can the time that a person takes to complete a marathon be predicted by knowing their stride length?

Variables: Stride length in cm is the explanatory (independent) variable and time to complete the marathon in minutes is the response (dependent) variable.

Trend: From the graph there is a strong negative linear relationship.

Association: People with longer stride lengths tend to finish the marathon in quicker times. This would be expected, because with each stride they cover a greater distance, so finish the marathon quicker.

Relationship: The relationship is strong and negative. This can be seen because the times to complete the marathon are fairly close to the line of best fit for almost all of the stride lengths and the line of best fit has a negative gradient. The correlation coefficient, being -0.9 supports that this is a strong relationship as it is close to -1, which would be a perfect straight line.

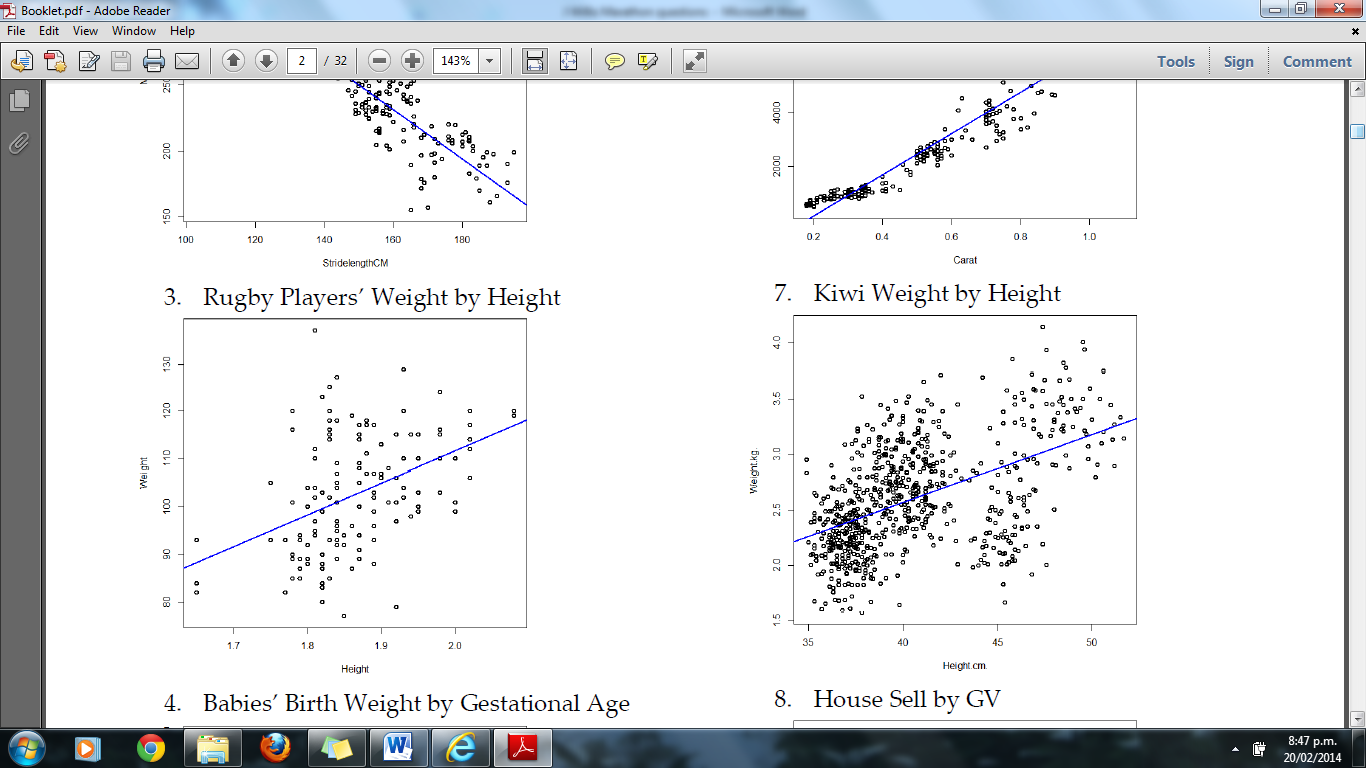
Scatter: The scatter about the regression line is fairly consistent throughout the stride lengths of between 100 cm and 195 cm. There is quite a lot of variation for runners with strides between 160 and 170 cm, however.

Outliers – there are no clear outliers, although the fastest times are run by runners with stride length about 165 cm and 170 cm which is about 20cm shorter than the longest stride.

Groups: There is no evidence of any grouping.

**3. Rugby**

|  |  |
| --- | --- |
| The data is real data and comes from http://www.rugby-sidestep-central.com/ **Variable** | **Description** |
| Country | New Zealand or South Africa |
| Position | Forward or Back |
| Weight | The weight of the player in kilograms (kg) |
| Height | The height of the player in metres (m) |



Summary of Rugby Players’ Weight by Height – Linear Trend

Weight = 66.726 \* Height + -21.87

Correlation = 0.43443

I wonder if there is a relationship between the height of a rugby player and their weight. Can knowing the height of a rugby player be used to predict their weight?

Variables: Height in m is the explanatory (independent) variable and weight in kg is the response (dependent) variable.

Trend: From the graph there is a very weak positive linear relationship.

Association: There is a slight tendency that taller rugby players are heavier, however this is not a strong association. In the general population, we would expect that taller people tend to have a greater volume and hence tend to weigh more than shorter people. In rugby, backs tend to have to run faster around the field so they may have more representatives who are taller than average, but leaner, meaning tall backs are not necessarily heavier. The forwards tend to be very solid players to pack the scrum and some backs can be quite short even though they are very heavy.

Relationship: The relationship is weak but positive. This can be seen because the weights are fairly scattered and many are a long way from the line of best fit for a given height and the line of best fit has a positive gradient. The correlation coefficient of 0.43 supports that this is a weak relationship, because a correlation between 0.4 and 0.6 is considered to be a weak relationship.

Scatter: The scatter about the regression line is reasonably close to the line of regression for the few rugby players with heights less than 1.75 m or greater than 2.0 m. In between there is considerable spread in the weights so that any linear pattern is not very clear. This would mean a prediction of weight given a rugby player’s height would not be expected to be very accurate for these heights.

Outliers – there are no clear outliers, although the heaviest rugby player is more than 20cm shorter than the tallest rugby player. (Note: The shortest 3 players may be being a strong influence on the position of the trend line and it would be interesting to investigate how influential they are by removing them, just to see. This does NOT mean that you can just remove values that might be outliers, but is just checking to see how influential they are)

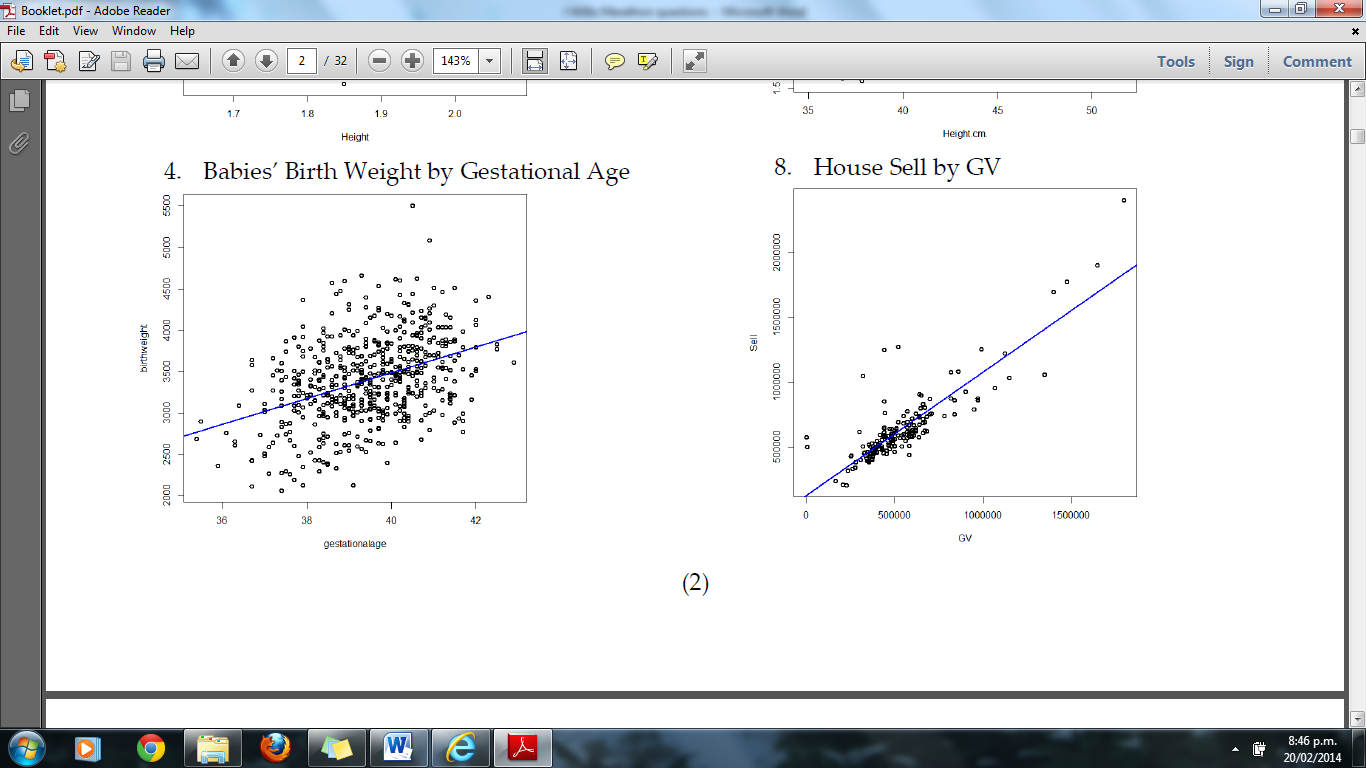
Groups: There is no evidence of any grouping.

**4. Babies**

Various health measures on new born babies and their mothers can give an indication of the future health of the infant. In particular, low birth weight is known to be associated with increased morbidity and poor health outcomes.

Data is routinely collected by all birthing centres in New Zealand concerning various health measures of mothers and their new born babies. A random sample of 550 records was selected in 2011 by a team of medical researchers from a birthing centre in a large teaching hospital.

|  |  |
| --- | --- |
| You have been supplied with the dataset containing some of the variables for the random sample collected in 2011. **Variable** | **Description** |
| bloodsugar | GDM = mother has gestational diabetes  Normal = mother has normal blood sugar levels |
| smoking | Smoker = mother smoked during pregnancy  NonSmoker = mother was a non-smoker during pregnancy |
| neonatalsexgroup | Male = new born infant is male  Female = new born infant is female |
| birthweight | Weight of infant at birth (in grams) |
| gestationalage | Length of pregnancy (in weeks) |
| fastingbloodglucose | Results from a routine blood test during pregnancy (mmol/L) |



Summary of Birth Weight by Gestational Age – Linear Trend

birthweight = 156.11 \* gestationalage + -2757.19

Correlation = 0.39553

I wonder if there is a relationship between the gestation period of a baby and their birth weight. Can knowing the time of gestation be used to predict a baby’s birthweight?

Variables: Gestation period in weeks is the explanatory (independent) variable and birth weight in g is the response (dependent) variable.

Trend: From the graph there is a very weak positive linear relationship.

Association: There is a slight tendency that babies who have a longer gestation period are heavier at birth. However, this is not a strong association. A baby who is born prematurely would not expect to be as heavy generally as a baby who has grown in the womb to full term (about 38 - 40 weeks) and if a baby is getting the nutritional requirements in the womb then the longer they are there, the more they would be expected to grow.

Retrieved from <http://www.sciencedaily.com/releases/2013/08/130806203327.htm>

“Normally, women are given a date for the likely delivery of their baby that is calculated as 280 days after the onset of their last menstrual period. Yet only four percent of women deliver at 280 days and only 70% deliver within 10 days of their estimated due date, even when the date is calculated with the help of ultrasound.

Now, for the first time, researchers in the USA have been able to pinpoint the precise point at which a woman ovulates and a fertilised embryo implants in the womb during a naturally conceived pregnancy, and follow the pregnancy through to delivery. Using this information, they have been able to calculate the length of 125 pregnancies.

"We found that the average time from ovulation to birth was 268 days -- 38 weeks and two days," said Dr Anne Marie Jukic, a postdoctoral fellow in the Epidemiology Branch at the National Institute of Environmental Health Sciences (Durham, USA), part of the National Institutes for Health. "However, even after we had excluded six pre-term births, we found that the length of the pregnancies varied by as much as 37 days”.”

Relationship: The relationship is weak but positive. This can be seen because the birth weights are fairly scattered and many are a long way from the line of best fit at a given gestation period. The line of best fit has a positive gradient. The correlation coefficient, being 0.40 supports that this is a very weak relationship because a correlation between 0.4 and 0.6 is considered to be a weak relationship.

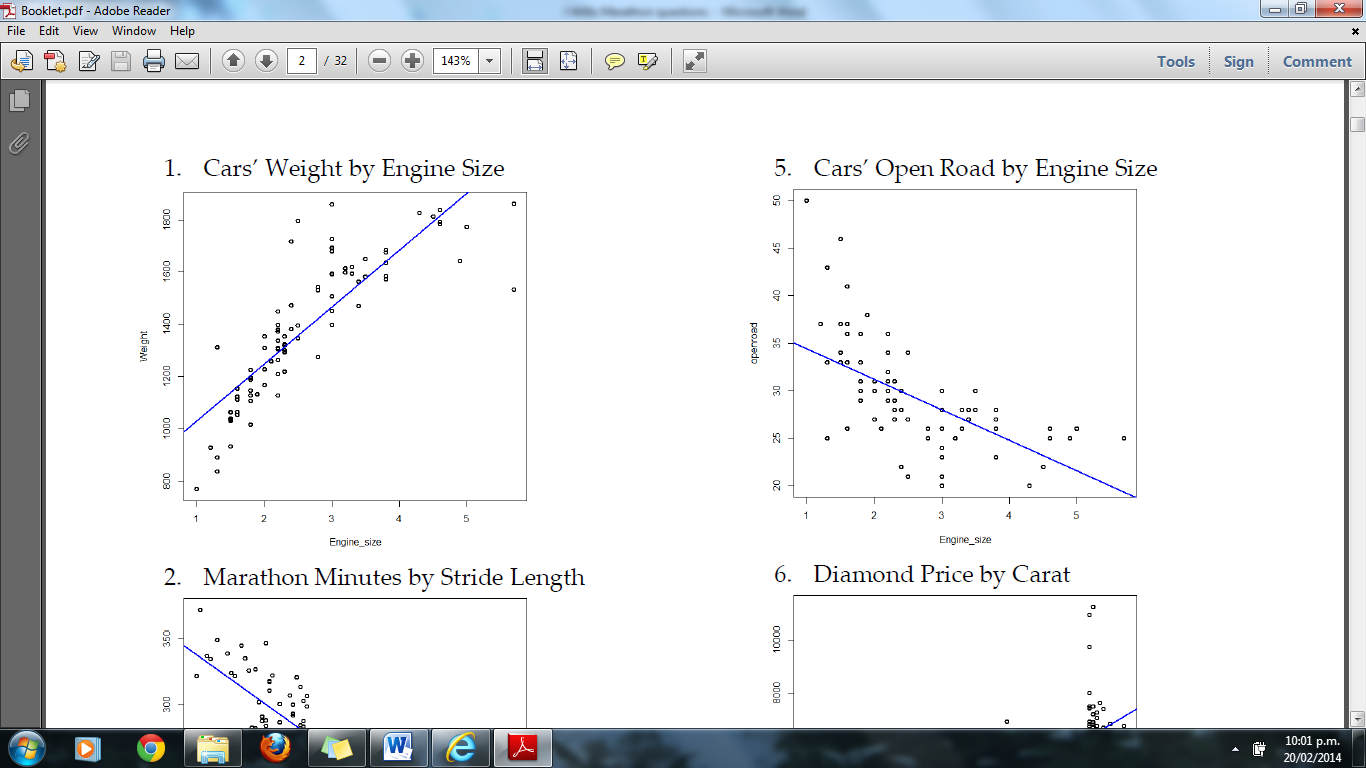
Scatter: The scatter about the regression line is reasonably consistent.

Outliers – there are no clear outliers, although there are 2 very heavy babies quite a lot heavier than the other more bunched group.

Groups: There is no evidence of any grouping.

**5. Cars**

|  |  |
| --- | --- |
| With rising costs of owning and running a car, and environmental awareness, buyers are becoming more conscious of the features when purchasing new cars. The data supplied is for new vehicles sold in America in 1993. **Variable** | **Description** |
| Vehicle Name | |
| Origin | Country of manufacture   America   Foreign |
| Price | US $1000 |
| Type | Small, midsize, large, compact, sporty, van |
| City | Fuel efficiency in kilometres per litre in cities and on motorways |
| OpenRoad | Fuel efficiency in kilometres per litre on country and open roads |
| Drive Train | Front Wheel Drive  Rear Wheel Drive |
| Engine Size | Size in litres |
| Manual Transmission | Yes  No |
| Weight | Weight of car in Kg |



Summary of Open Road by Engine Size – Linear Trend

openroad = -3.2215 \* Engine\_size + 37.68

Correlation = -0.62679

I wonder if there is a relationship between the engine size of a car and the fuel efficiency of a car travelling on the open road. Can the amount of km that a car will run on a litre of petrol be predicted by knowing the car’s engine size?

Variables: Engine size is the explanatory (independent) variable and number of km a car will run on the open road per litre of fuel is the response (dependent) variable.

Trend: From the graph there is a moderate negative linear relationship. However, because the fuel efficiency is generally above the trend line for smaller engine cars and then may be above for large engine sizes (and it makes no sense to get no km or negative km per litre), a better relationship may be a curve (similar to a hyperbola such as y = - but this is beyond the scope of this course to use this model).

Association: Cars with larger engine sizes tend to travel shorter distance per litre of fuel. This would be expected, because the more powerful cars would be expected to use more petrol in order to produce this power.

Relationship: The relationship is moderate negative and follows a general linear trend. This can be seen because the fuel efficiency is fairly scattered from the line at any given engine size. There is a negative trend shown by the line of best fit having a negative gradient. The correlation coefficient, being -0.6 supports that this is a moderate to weak relationship.

Scatter: The scatter about the regression line is greatest for cars with small engine size where there are some highly efficient cars (points well above the trend line) but also some very poor performances. There will be other factors such as the age of the car, the mechanical fitness of the car which may influence how fuel efficient these cars are.

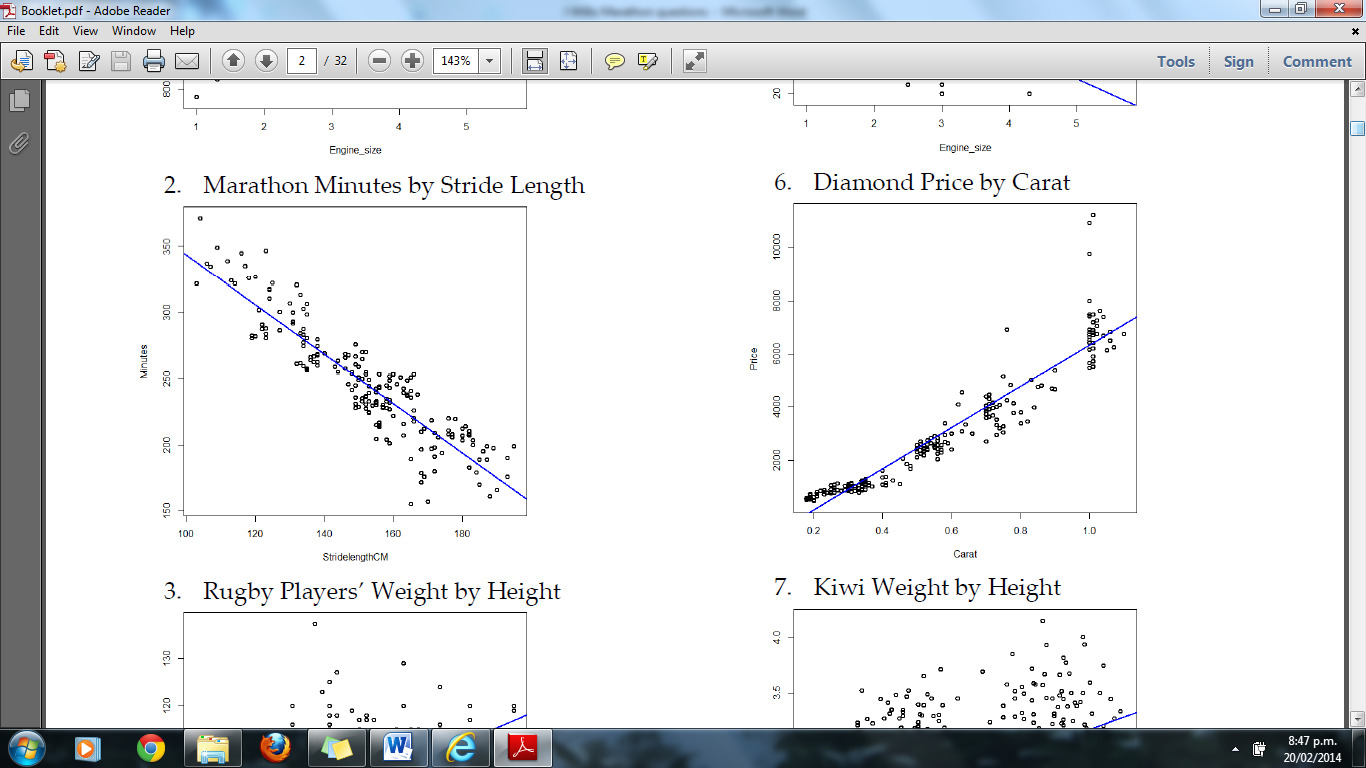
Outliers – there are no clear outliers (Note: while the highly efficient small cars may be thought of as outliers, they are all reasonably close to each other and other cars, so they are not considered to be outliers)

Groups: There is no evidence of any grouping.

**6. Diamonds**

Every diamond is unique, and there are a variety of factors which affect the price of a diamond. Insurance companies in particular are concerned that stones are valued correctly.

|  |  |
| --- | --- |
| Data on 308 round diamond stones was collected from a Singapore based retailer of diamond jewellery, who had the stones valued. **Variable** | **Description** |
| Carat | Weight of diamond stones in carat units 1 carat = 0.2 grams |
| Colour | Numerical value given for quality of colour ranging from 1=colourless to 6=near colourless |
| Clarity | Average = score 1, 2 or 3  Above average = score 4, 5 or 6 |
| Lab | Laboratory that tested & valued the diamond  1 = laboratory 1  2 = laboratory 2 |
| Price | Price in US dollars |



Summary of Diamond Price by Carat – Linear Trend

Price = 7788.7 \* Carat + -1450.02

Correlation = 0.94701

I wonder if there is a relationship between the carat measurement in diamonds and the price of diamonds in order to predict the price of a diamond by its carat measure.

Variables: Number of carats is the explanatory (independent) variable and price in American dollars is the response (dependent) variable.

Trend: From the graph there is a strong positive linear relationship.

Association: Diamonds with higher carat weight tend to be more expensive. This is expected since heavier diamonds will contain more of the material and therefore be more expensive.

Relationship: The relationship is strong because most of the diamonds have their prices quite close to the line of best fit for any carat value. The line of best fit has a positive gradient. The correlation coefficient, being 0.95 supports that this is a strong relationship as this is close to r = 1 which would be a perfect straight line.

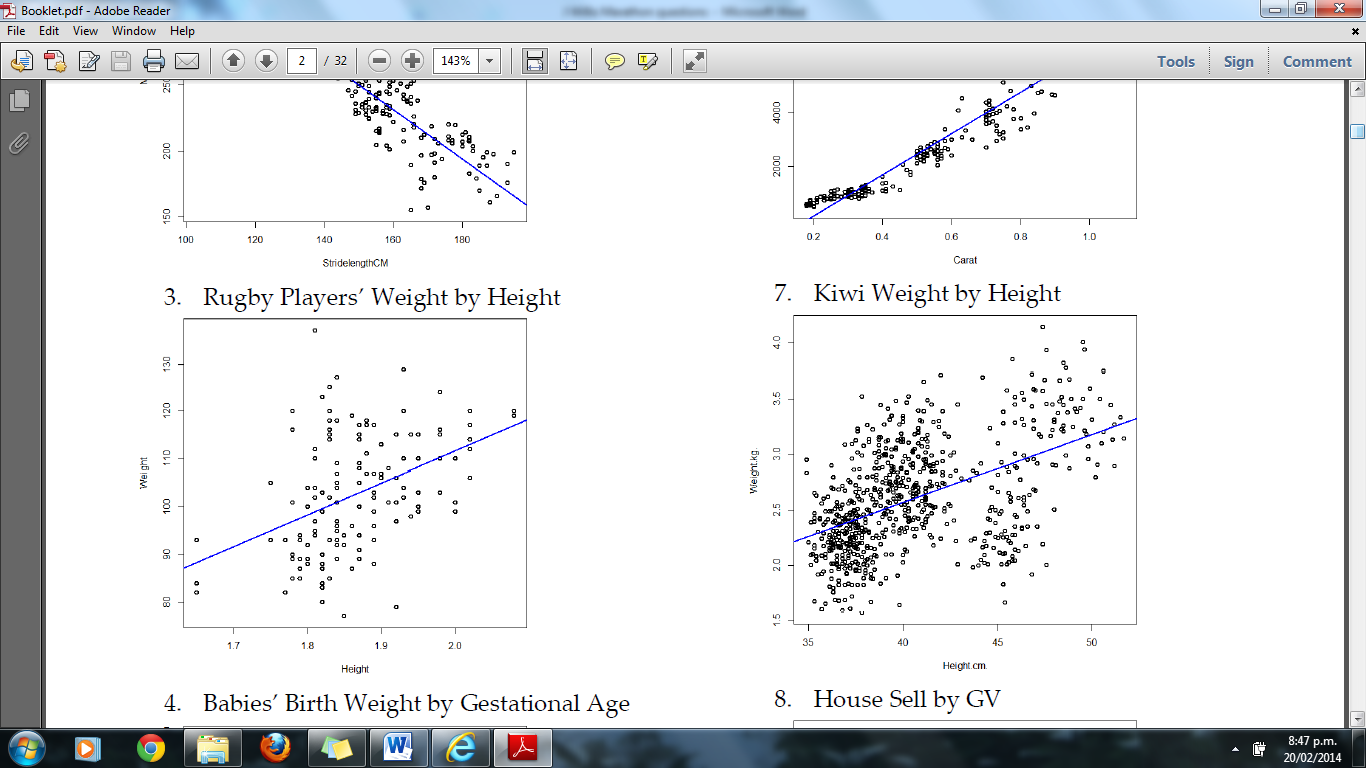
Scatter: The scatter about the regression line is greater for diamonds with smaller carat value and gets more spread as the carat weight increases. This may suggest that a curve model may be appropriate as well. There is greater variation for diamonds with 1 carat that any others, which may indicate that at this size other factors such as clarity may become more influential in determining the price.

Outliers – there are no clear outliers although three of the diamonds about 1 carat are much more expensive than the others.

Groups: There is some evidence of grouping with a bunch less than 0.4 carats, a bunch around 0.55 carats, 0.75 carats and 1 carat. All of these bunches still fit the trend line reasonably closely.

**7. Kiwi**

|  |  |
| --- | --- |
| Data on kiwi birds around New Zealand was collected in order to help with conservation efforts. **Variable** | **Description** |
| Species | GS-Great Spotted  NIBr-NorthIsland Brown  Tok-Southern Tokoeka |
| Gender | M-Male  F-Female |
| Weight(kg) | The weight of the kiwi bird in kg |
| Height(cm) | The height of the kiwi bird in cm |
| Location | NWN-North West Nelson  CW-Central Westland  EC-Eastern Canterbury  StI-Stewart Island  NF-North Fiordland  SF-South Fiordland  N-Northland  E-East North Island  W-West North Island |



7. Summary of Kiwi Weight by Height – Linear Trend

Weight.kg. = 0.061319 \* Height.cm. + 0.12

Correlation = 0.50383

I wonder if there is a relationship between the height of a kiwi and the weight of a kiwi, in order to predict a kiwi’s weight by its height. This might be used to determine whether a kiwi is in good health or not – if a kiwi’s weight is well below the line of best fit, this could indicate under-nourishment or if it is well above it could indicate another problem (do kiwis get eating disorders?).

Variables: Height in cm of the kiwi is the explanatory (independent) variable and weight in kg is the response (dependent) variable.

Trend: From the graph there is a weak to moderate positive linear relationship.

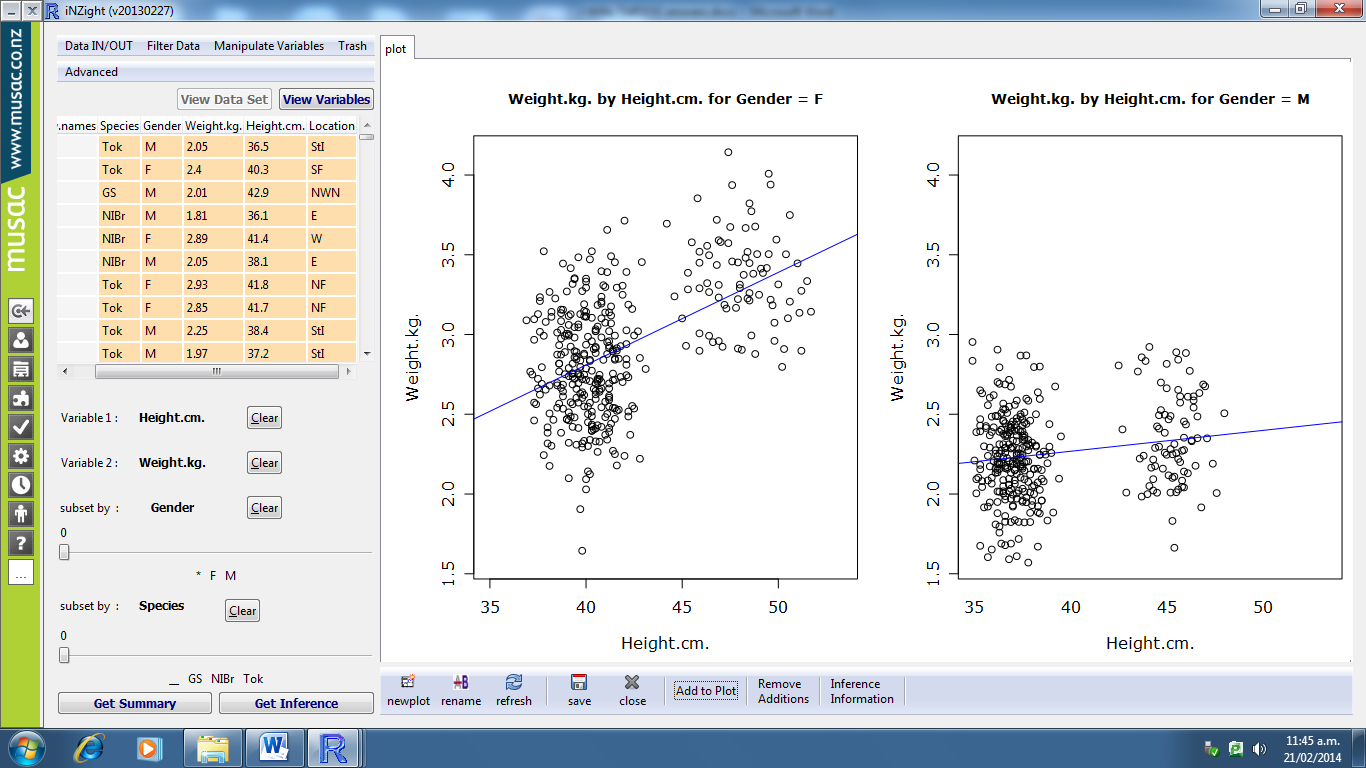
Association: There is a slight tendency for taller kiwis to be associated with heavier kiwis, although this is generally a fairly weak association. It might be expected since taller kiwis would tend to carry more bulk and hence be heavier.

Relationship: The relationship is weak because most of the kiwi weights are not very close to the line of best fit for any given height of that kiwi. The line of best fit has a positive gradient. The correlation coefficient, being 0.5 supports that this is a weak relationship as r values of between 0.4 and 0.6 tend to indicate a weak relationship.

Scatter: The scatter about the regression line is greater for diamonds with smaller carat value and gets more spread as the carat weight increases. This may suggest that a curve model may be appropriate as well. There is greater variation for diamonds with 1 carat that any others, which may indicate that at this size other factors such as clarity may become more influential in determining the price.

Outliers – there are no clear outliers.

Groups: There is strong evidence of grouping with one bunch of kiwis with heights between 35 and 42 cm. There is another group with heights above 45 cm and this could possibly be split between those less than 2.7 kg in weight and those more than 1 kg in weight. In fact this warrants further investigation and it can be determined that there is a distinction between males and females with females being heavier. There is also a distinct grouping with the Great Spotted species being taller than the other kiwis. (While this is getting into Merit / Excellence, to not recognise these groupings would mean a student would be unlikely to get Merit)

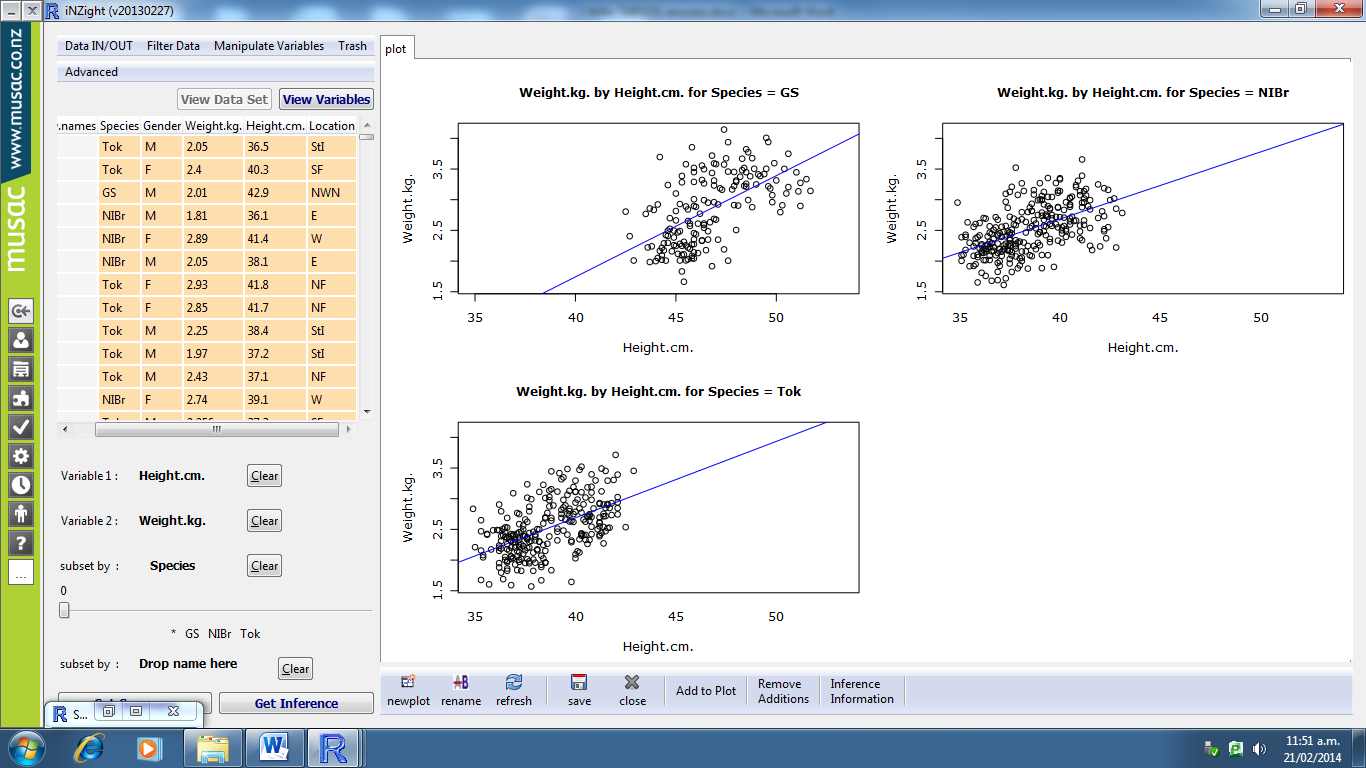


Summary for Gender = F Summary for Gender = M

Linear Trend Linear Trend

Weight.kg. = 0.058068 \* Height.cm. + 0.4862 Weight.kg. = 0.013078 \* Height.cm. + 1.7448

Correlation = 0.5251 Correlation = 0.17691



Summary for Species = GS Summary for Species = NIBr Summary for Species = Tok

Linear Trend Linear Trend Linear Trend

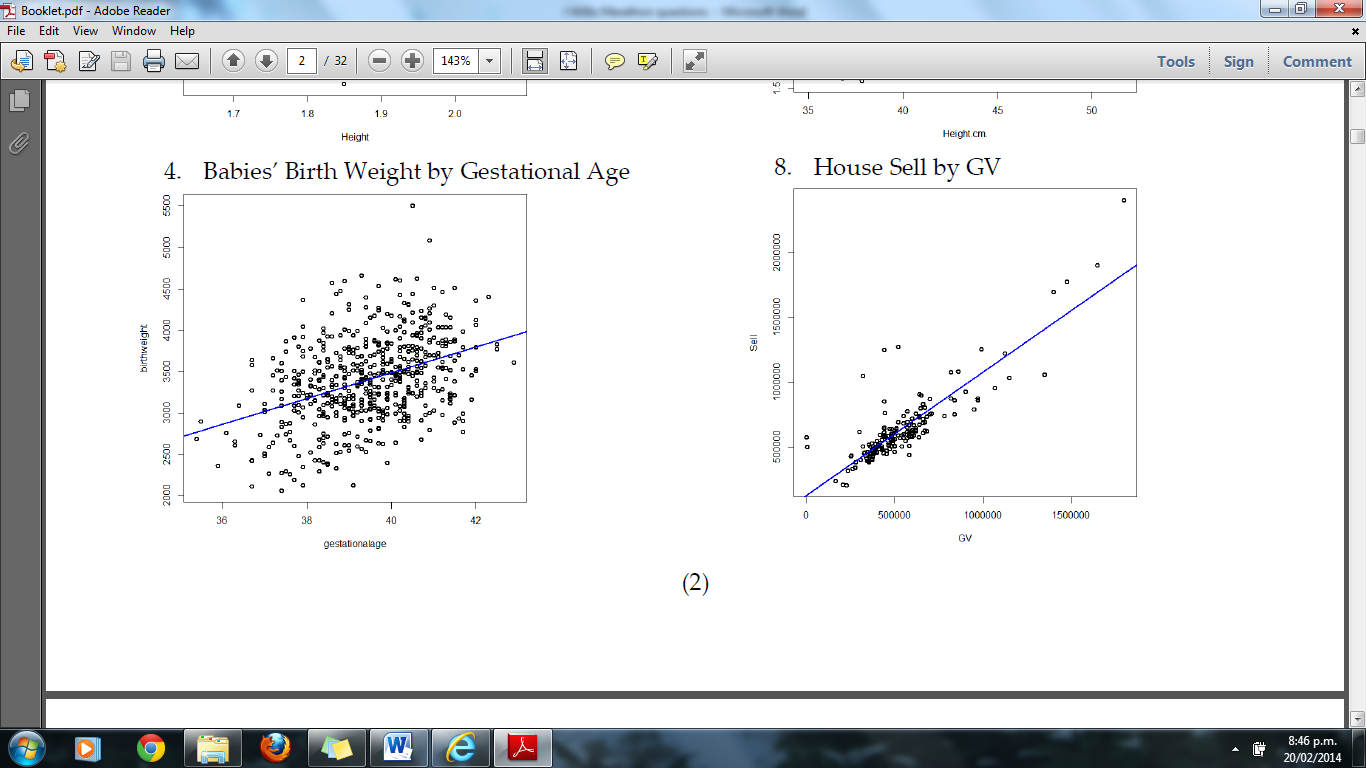
Weight.kg. = 0.16495 \* Height.cm. + -4.8557 Weight.kg. = 0.10946 \* Height.cm. + -1.693 Weight.kg. = 0.12444 \* Height.cm. + -2.2881

Correlation = 0.57989 Correlation = 0.52905 Correlation = 0.53939

**8. NorthShoreHouses**

When buying houses a number of factors are taken into considered including location, type of house, number of bedrooms and the size of section.

|  |  |
| --- | --- |
| The data set supplied has information from recorded house sales on the North Shore during 2007 & 2008 **Variable** | **Description** |
| Date | Year and month e.g. 200801 is January 2008 |
| Suburb | Browns Bay  Murrays/Mairangi Bay |
| List | Listing price |
| Sell | Selling price |
| Days | The number of days the house was on the market |
| Bedroom | Number of Bedrooms |
| Type | R = Residential House  U = Townhouse or Unit |
| Land | Area of section in square metres |
| GV | Government valuation $ |



8. Summary of Sell by GV

Sell = 0.94873 \* GV + 128812.86

Correlation = 0.85765

I wonder if there is a relationship between the government valuation of a house and the selling price of houses sold in North Shore.

Variables: Government is the explanatory (independent) variable and selling price is the response (dependent) variable, both in $NZ.

Trend: From the graph there is a strong positive linear relationship.

Association: The government valuation is found for rating purposes and is often based on selling prices of houses in a particular area so is likely to be a good indicator of the selling price of a house. Therefore the association of high government valuations with high selling prices would be expected.

Relationship: The relationship is strong because most of the selling prices are quite close to the line of best fit for any government valuation. The line of best fit has a positive gradient. The correlation coefficient, being 0.86 supports that this is quite a strong relationship as this is reasonably close to r = 1 which would be a perfect straight line.

Scatter: The scatter about the regression line is consistent for all government valuations.

Outliers – there are no clear outliers.

Groups: There is no evidence of grouping, but there are many houses valued around $500 000 and few over $1 000 000. This data was collected in 2007 / 2008 so is likely to have changed as house price high costs have been the subject of concern in recent years.