# TODAY'S AGENDA

**01** Your modelling assignment

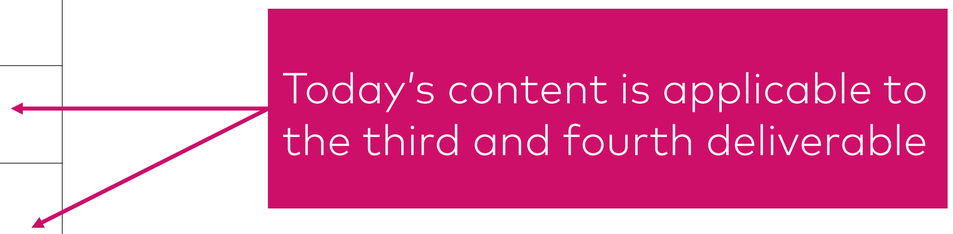
**02** Regression – predicting an event

**03** Building Tableau dashboards

**04** A little competition

# DELIVERABLES

| Deliverable | %Grade | Due Date |
|---|---|---|
| Capstone Deliverable 1: Capstone Proposal | 15% | September 30th |
| Capstone Deliverable 2: Cleaning and Visualization Report | 20% | October 21ST |
| Capstone Deliverable 3: Modeling | 15% | November 11th |
| Capstone Deliverable 4: Tableau Dashboard | 15% | November 25th |
| Capstone Final Deliverable 5: Presentation | 15% | December 2nd |
| Capstone Final Deliverable 6: Report | 20% | December 16th |
| Total | 100% | |



# DUE NEXT

- The third report “Modelling” is due on November 11th (6 days away!)
- This deliverable is where I see the creative juices flowing, I look forward to it!
- Rubric is on the portal

# TODAY'S CLASS

## AND HOW IT RELATES TO ASSIGNMENT 3

# FOR YOUR ASSIGNMENT 3

| Criteria | Description | Deduction | Overall Marks |
|---|---|---|---|
| Predictive modelling technique and evaluation | Correct use of a modelling algorithm.<br><br>Evaluation of the model and a description on the "accuracy" metric you chose. Why was this chosen? What else did you look at? | Technical problems with the algorithm | /30 |
| Predictive modelling solution addresses the business problem | Business explanation on why a modelling solution is applicable to this project and the ideal solution. | Section is missing completely | /15 |
| | For example, we predict when someone is likely to unsubscribe from a newsletter as we are predicting when a NEW unsubscribe will happen.<br>We predict the attendance for a NEW movie in the future, given the movie/environment.<br><br>It is easy to solve your problem with descriptive statistics – this is a trend analysis.<br><br>Make sure your problem needed a predictive modelling solution and explain the benefit. | | |
| Code presentation | Your script is easy to read and understand.<br><br>If you were to put this code into production, other analysts will be able to follow your thought process. | The submission is a wall of text with no comments or sectioning to structure the flow of your code<br><br>No code was submitted | /10 |
| Overall deductions | Overall marks will be deducted for:<br>• Report lacks professionalism and quality<br>• Report lacks a business perspective | | |

# PREDICTING A NUMBER

Last class we went through predicting an event (Classification). Today we will predict a number.

There are many ways to evaluate your regression, let's go through these three:

1. R Squared/adjusted R-squared
2. Root mean square error (RMSE)
3. Mean Absolute Error (MAE)

# R-SQUARED

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

**Definition:** R Squared measures how much of variability of the dependent variable can be explained by the model. It is square of Correlation Coefficient(R)

**Is my R-squared good?:** The range for R squared if from 0 to 1. Bigger value means better.

**Pros/Cons:**

- Nice, interpretable value back to the business
- Prone to overfitting problems. If you model is overfit, R-squared will not take this into consideration. You can use Adjusted R-squared which penalizes your model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

# RMSE

**Definition:** Mean Square Error is an absolute measure of the goodness for the fit. It gives you an absolute number on how much your predicted results deviate from the actual number.

**Is my RMSE good?:** The RMSE is in the units of your target. I usually look at the full range that my target is spread over. If I'm predicting the price of a Cineplex concessions spend, the range is likely {$1.00 - $26.00}. If my RMSE is $20,000, then it is poor. If my RMSE is $2.00 this is acceptable!

**Pros/Cons:**

- RMSE is in the units of your target so it is very interpretable. For example, you're predicting sales and the RMSE = $50, means your model is off by $50
- Emphasizes greater errors than smaller ones (due to the square function)

# MAE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

**Definition:** Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error.

**Is my MAE good?:** The range for R squared if from 0 to 1. Bigger value means better.

**Pros/Cons:**

- MAE does not have a squared function, thus it isn't penalizing a big error like MSE
- It is a nice average term, that summarize what your average error is (nice and explanatory!)
- Business people might not truly get if it's good or not

# APPLIED EXAMPLE

Korite is a jewellery store. They sell fine gems and jewellery, but their main source of income is their diamond sales.

Pricing diamonds manually is a tedious process. It relies on a few different elements (such as colour, cut, clarity, polish, symmetry).

Korite wants us to create a model to price new diamonds coming into their store.

# REMEMBER OUR PROCESS SO FAR

1. ✓ Describe the problem statement (M1)
2. ✓ Clean the data, remove outliers and engineer new feature (M3)
3. ✓ Explore the data and answer preliminary questions (EDA) (M3/M4)
4. → Build a model to answer your problem statement ←

## Step 1: The problem statement

Korite is a chain of jewellery stores that sells fine jewellery, including diamonds.

The process to price diamonds is very time consuming, but pricing is a necessary task to make sure the final piece of jewellery is sold at the price it is worth.

An intelligent algorithm such as a regression that predicts the price of a diamond, will benefit the staff as less time will be spent manually pricing diamonds, and it will benefit the overall company as it will ensure accurate pricing standards.

Additionally, we can utilize this system to understand what features of a diamond make it worth more.

Our goal is to lift sales by 1% incrementally.

https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/in-ra-analytics-in-the-jewellery-industry-noexp.pdf

# LOOK AT THE DATA

| ID | Carat Weight | Cut | Color | Clarity | Polish | Symmetry | Report | Price |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.1 | Ideal | H | SI1 | VG | EX | GIA | 5169 |
| 2 | 0.83 | Ideal | H | VS1 | ID | ID | AGSL | 3470 |
| 3 | 0.85 | Ideal | H | SI1 | EX | EX | GIA | 3183 |
| 4 | 0.91 | Ideal | E | SI1 | VG | VG | GIA | 4370 |
| 5 | 0.83 | Ideal | G | SI1 | EX | EX | GIA | 3171 |
| 6 | 1.53 | Ideal | E | SI1 | ID | ID | AGSL | 12791 |
| 7 | 1 | Very Good | D | SI1 | VG | G | GIA | 5747 |
| 8 | 1.5 | Fair | F | SI1 | VG | VG | GIA | 10450 |
| 9 | 2.11 | Ideal | H | SI1 | VG | VG | GIA | 18609 |
| 10 | 1.05 | Very Good | E | VS1 | VG | G | GIA | 7666 |

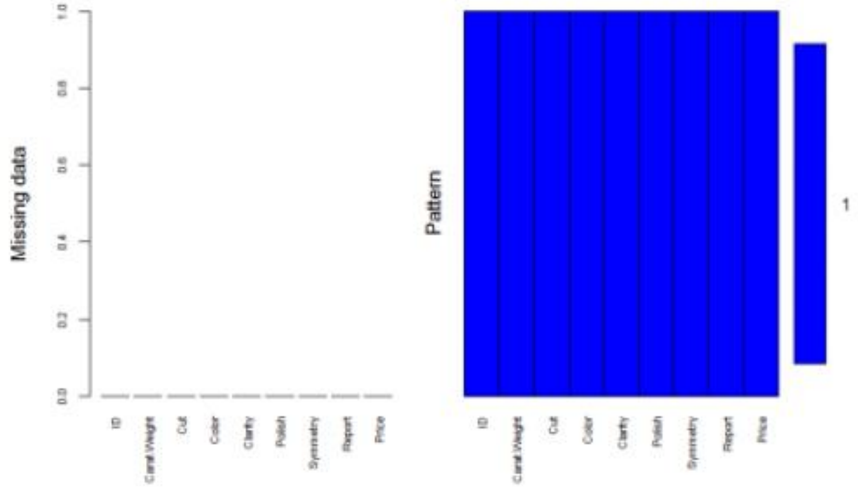
**Definitions:**

- Id: A unique identifier for each row.
  - Each row represents 1 diamond
  - From ID 6001 and higher, there are no prices. This is what you are predicting
- Carat Weight: The measure of physical weight of a diamond. 1 ct = .2 grams
- Cut: A measure of the diamond's proportions
- Color: Diamonds come in a variety of colours. D = colourless (better)... I = near colorless (worse)
- Clarity: Grade for the visual appearance of a diamond. IF = Internally Flawless
- Polish: How well light can pass through the diamond
- Symmetry: Degree of smoothness
- Report: Assessment of the 4C's of a diamond
- Price: Our response variable in dollars

## Step 2: Clean missing values (if they exist)

**Key Finding:** No data is missing in this dataset

**Missing value analysis:**

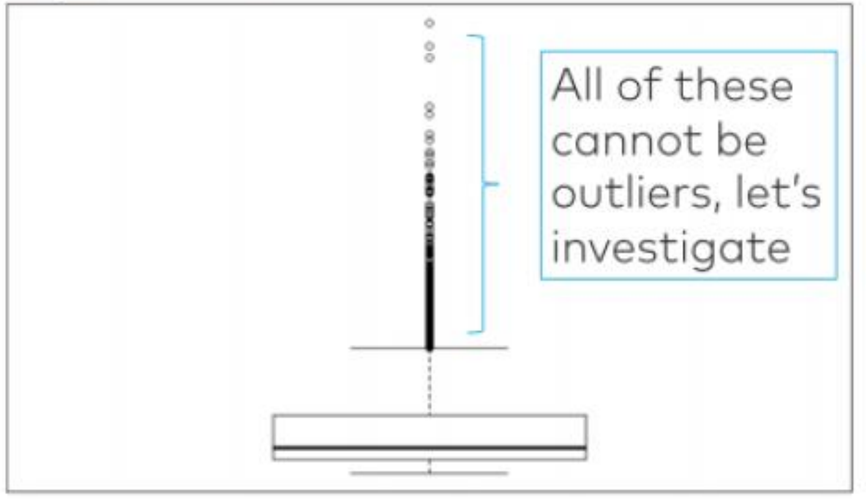



```
aggr(dataset, col=c('blue','red'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(dataset), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```

## Seek out outliers

**Key Finding:** Data looks suspicious and needs more investigation. This graph is illustrating that a significant chunk of the dataset are outliers. An inference is that the data is skewed, and there are possibly price values which are very high

**Boxplot of Price**

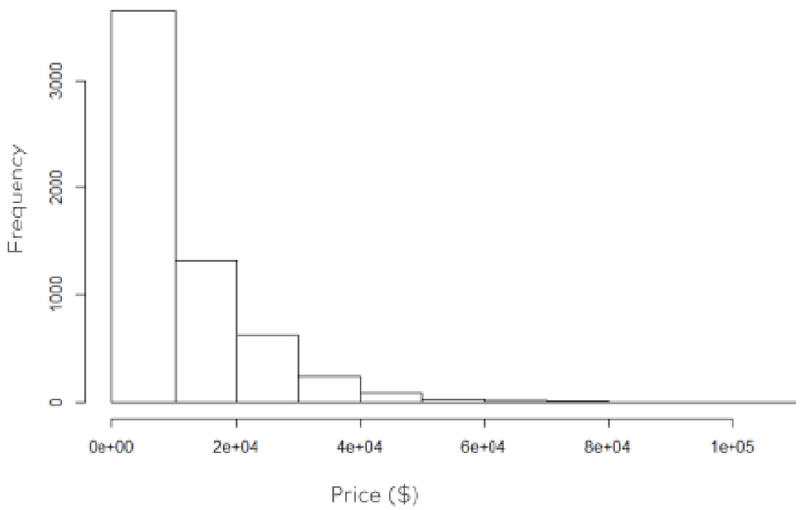



```
boxplot(diamond.data.training$Price)
```

## Outlier investigation

**Key Finding:** Data it NOT normal and is skewed
We see that the data has a very long left tail, suggesting that there are some diamond prices that are very high. This is likely the nature of the data. We will not delete these data points, but instead correct the skewness of the graph

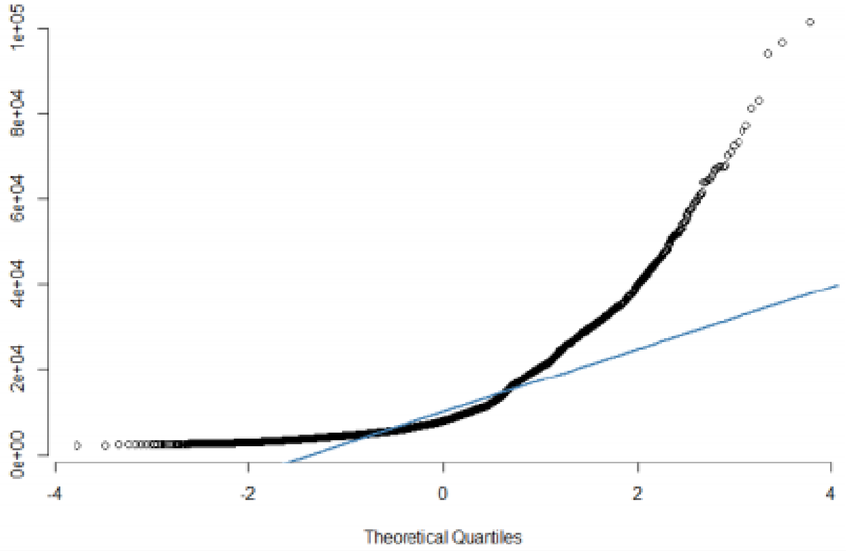
**Total number of diamonds by Price ($)**




```
hist(diamond.data.training$Price)
```

**Key Finding:** The qq plot is an additional graph that illustrated the data is not normal. We plotted this to reinforce our assumption that the data is skewed

**QQ Plot**

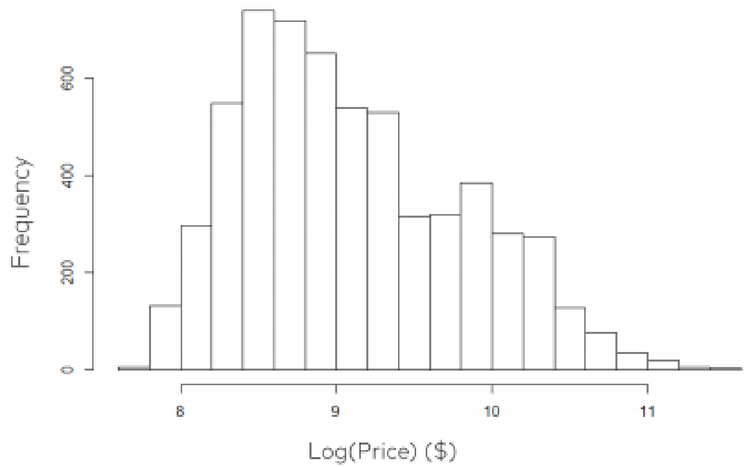



```
qqnorm(diamond.data.training$Price, pch = 1, frame = FALSE)
qqline(diamond.data.training$Price, col = "steelblue", lwd = 2)
```

## (continued): Outlier investigation & transformation

**Key Finding:** Skewness was corrected with a log transformation
The resulting shape of Price is balanced and will be able to explain more variance in the model

**Total number of diamonds by Log(Price) ($)**




```
hist(log(diamond.data.training$Price))
```

**In Summary:**

- There are prices which are very high
- I am choosing to keep these in my model on purpose as I believe they are representative of reality
- However, I will apply a transformation to prepare the data for my predictive model, and correct the skewness which occurs due to high prices

**Step 3: EDA**

**Remember: Always centre your thoughts to your response variable**


Some ideas I have:

1. Bigger (carat weight) diamonds should **cost** more
2. Clearer diamonds seem as thought they will be more **expensive**
3. There is a marketing "appeal" by saying your diamond is 1 carat or larger, which will lead to more people **buying big diamonds**

**Let's go to Tableau**!

# TEST & TRAIN

```
dataset <-read.csv(file.choose(), header=TRUE, sep=",")
diamond.data.training    <-subset(dataset, ID<=6000)  #train
diamond.data.prediction  <-subset(dataset, ID>=6001)  #test
```

This data set can be easily split into Test and Train as the latter entries are what we need to populate

## Step 4: Regression (BASIC)

Create a basic model with all the features

```
fit<-lm(Price~., data=diamond.data.training)
summary(fit)
```

Examine results

We see:

- Carat weight is significant
- Cololur is significant
- Polish, Symmetry and Report not too much

```
Call:
lm(formula = Price ~ ., data = diamond.data.training)

Residuals:
   Min     1Q Median     3Q    Max
-21704  -2105    -400   1532  50446

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.632e+04  1.956e+03  13.455  < 2e-16 ***
ID                3.720e-02  2.813e-02   1.322 0.186125
Carat.Weight      1.839e+04  1.051e+02 175.010  < 2e-16 ***
CutGood          -3.272e+02  3.625e+02  -0.903 0.366742
CutIdeal          2.692e+02  3.554e+02   0.757 0.448859
CutSignature-Ideal 1.667e+03  4.398e+02   3.791 0.000152 ***
CutVery Good     -3.994e+01  3.455e+02  -0.116 0.907985
ColorE           -2.329e+03  2.005e+02 -11.620  < 2e-16 ***
ColorF           -3.083e+03  1.898e+02 -16.239  < 2e-16 ***
ColorG           -4.804e+03  1.783e+02 -26.944  < 2e-16 ***
ColorH           -6.365e+03  1.881e+02 -33.842  < 2e-16 ***
ColorI           -8.042e+03  1.928e+02 -41.709  < 2e-16 ***
ClarityIF        -2.710e+04  1.912e+03 -14.173  < 2e-16 ***
ClaritySI1       -3.690e+04  1.898e+03 -19.438  < 2e-16 ***
ClarityVS1       -3.397e+04  1.899e+03 -17.889  < 2e-16 ***
ClarityVS2       -3.532e+04  1.899e+03 -18.599  < 2e-16 ***
ClarityVVS1      -3.058e+04  1.909e+03 -16.016  < 2e-16 ***
ClarityVVS2      -3.243e+04  1.902e+03 -17.054  < 2e-16 ***
PolishG           5.585e+00  2.018e+02   0.028 0.977918
PolishID         -5.676e+02  7.450e+02  -0.762 0.446174
PolishVG         -1.564e+02  1.266e+02  -1.236 0.216645
SymmetryG        -4.274e+02  1.889e+02  -2.262 0.023723 *
SymmetryID        1.386e+02  7.724e+02   0.179 0.857643
SymmetryVG       -2.929e+02  1.336e+02  -2.193 0.028352 *
ReportGIA         1.858e+02  3.475e+02   0.535 0.592897
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3768 on 5975 degrees of freedom
Multiple R-squared:  0.8636,    Adjusted R-squared:  0.8631
F-statistic:  1577 on 24 and 5975 DF,  p-value: < 2.2e-16
```

## Step 4 (continued): Regression (A bit better)

Build a log model (per our findings)

```
fit.log<-lm(log(Price)~(Carat.Weight)+Cut+Color+Clarity+Polish+Symmetry+Report, data=diamond.data.training)
summary(fit.log)
```

Examine results

We see:

- Adjusted R^2 has increased
- Only Symmetry is not significant

```
Call:
lm(formula = log(Price) ~ Carat.Weight + Cut + Color + Clarity +
    Polish + Symmetry + Report, data = diamond.data.training)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63531 -0.07345  0.02589  0.09211  0.56341

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.013732   0.074342 107.796  < 2e-16 ***
Carat.Weight      1.361388   0.004350 312.952  < 2e-16 ***
CutGood           0.060545   0.015039   4.026 5.76e-05 ***
CutIdeal          0.119900   0.014773   8.116 6.01e-16 ***
CutSignature-Ideal 0.268915  0.018235  14.747  < 2e-16 ***
CutVery Good      0.086276   0.014350   6.012 1.96e-09 ***
ColorE           -0.088815   0.008280 -10.726  < 2e-16 ***
ColorF           -0.110613   0.007823 -14.139  < 2e-16 ***
ColorG           -0.199272   0.007320 -27.222  < 2e-16 ***
ColorH           -0.329916   0.007734 -42.659  < 2e-16 ***
ColorI           -0.465346   0.007922 -58.740  < 2e-16 ***
ClarityIF        -0.277022   0.072475  -3.822 0.000134 ***
ClaritySI1       -0.863616   0.071836 -12.022  < 2e-16 ***
ClarityVS1       -0.611590   0.071876  -8.509  < 2e-16 ***
ClarityVS2       -0.697952   0.071859  -9.713  < 2e-16 ***
ClarityVVS1      -0.402974   0.072329  -5.571 2.66e-08 ***
ClarityVVS2      -0.470795   0.071986  -6.540 6.77e-11 ***
PolishG          -0.039152   0.008305  -4.714 2.49e-06 ***
PolishID          0.090800   0.032670   2.779 0.005468 **
PolishVG         -0.026565   0.005235  -5.075 4.02e-07 ***
SymmetryG        -0.011680   0.007780  -1.501 0.133353
SymmetryID       -0.010993   0.033568  -0.327 0.743319
SymmetryVG       -0.012573   0.005545  -2.268 0.023392 *
ReportGIA         0.086203   0.014290   6.033 1.73e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.[illegible]
Multiple R-squared:  0.9597,    Adjusted R-squared:  0.9596
F-statistic:  5158 on 23 and [illegible]
```

# THE MODEL

```
dataset <-read.csv(file.choose(), header=TRUE, sep=",")
diamond.data.training    <-subset(dataset, ID<=6000)  #train
diamond.data.prediction  <-subset(dataset, ID>=6001)  #test

fit<-lm(Price~., data=diamond.data.training)
predicted.prices<-predict(fit, diamond.data.prediction)
write.csv(predicted.prices, file = "Predicted Diamond Prices.csv")
```

Did you realize the model was 5 lines of code!

The beauty in models is in feature engineering, optimization and ensembles.

Creating a prediction is not difficult, don't be afraid!

# YOUR IMPACT

- Any analyst can run these models
- You are better than these analysts as you ~~attended my course~~ know the power of the "what does it mean"
- Be influential with your results and follow up with commentary on how your results can influence the business (or social cause)

- Results reveal that we can successfully predict the price of diamonds (adjusted R^2 = .96)
- By building this model into an interactive UI, staff can enter in the parameters of the diamond , which will generate pricing, saving time and reducing manual errors
- Additionally, we see that carat weight explains the price of the diamond the most
- We recommend that marketing campaigns and sales tactics enforce the impact of the "bigger diamond" in order to get more sales revenue to Michael Hill

# CREATE A TABLEAU PUBLIC LOGIN

Different to your Tableau desktop login. This allows you to be part of the online community

https://public.tableau.com/en-us/s/

# TABLEAU DASHBOARDS

# TABLEAU DASHBOARD

- A dashboard is a collection of several views, letting you compare a variety of data simultaneously

- Like worksheets, you access dashboards from tabs at the bottom of a workbook. Data in sheets and dashboards is connected; when you modify a sheet, any dashboards containing it change, and vice versa.

- TL/DR: A dashboard is a single view that has multiple worksheets (graphs) on it

https://www.artplusdata.com/

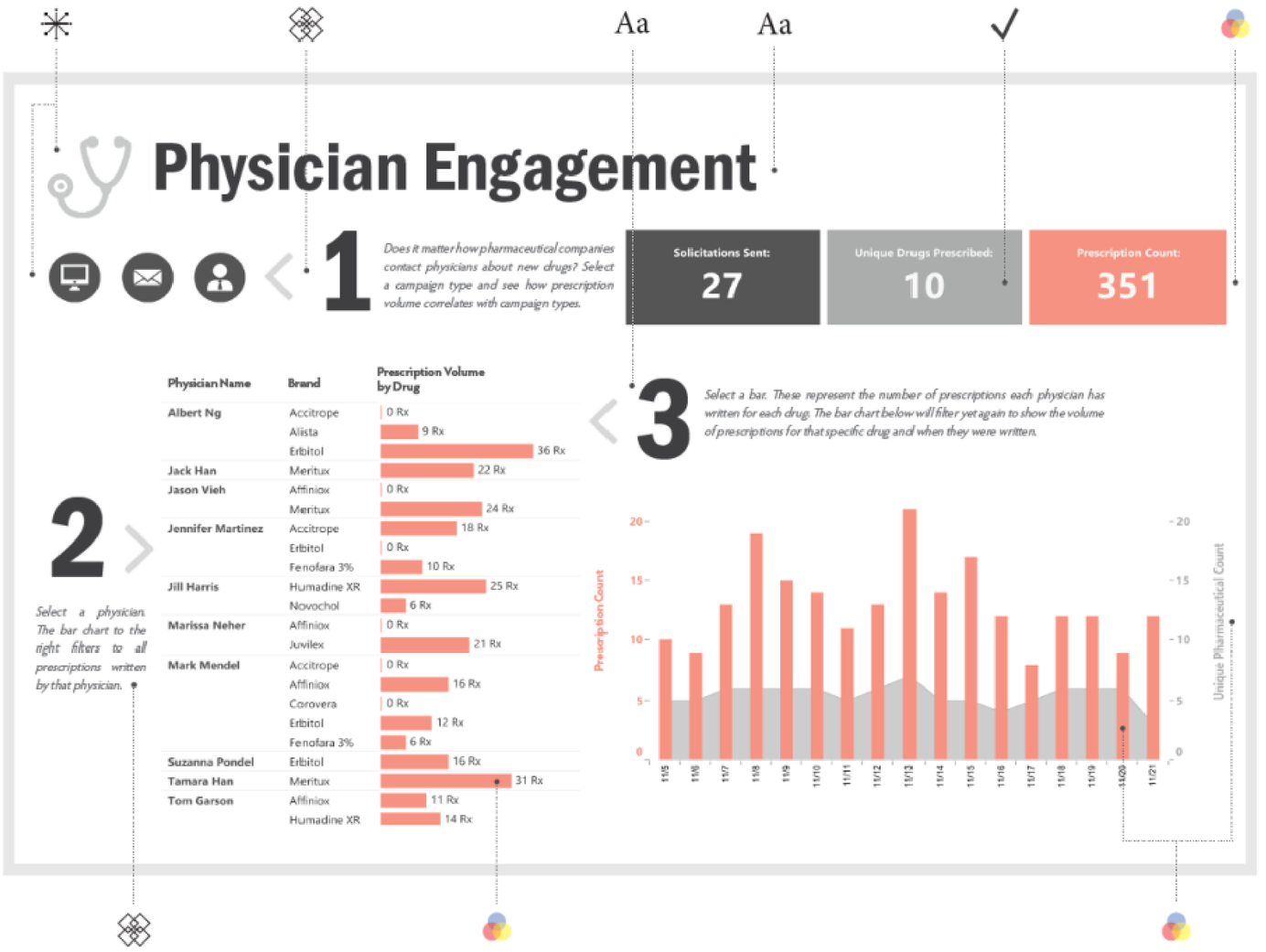
# ELEMENTS OF DASHBOARD DESIGN

1. Integrity: The dashboard needs to be powered by fairness. Pull in all relevant metrics to tell a fair story (do not be biased)

2. Flow: Have visual cues so the user can easily identify filters and move logically through the dashboard. Flow creates a technique called momentum. Users explore and create new ideas.

3. Colour: Use distinctive colours to point out special items. Less is more!

4. Charm: Adding charm is not always necessary but it adds a lot to usability and giving the impression that your dashboard is professional




Integrity (we have to trust)

Flow: Very clear with numbers

Colour: Easy to read palette

Charm: Icons and clickability

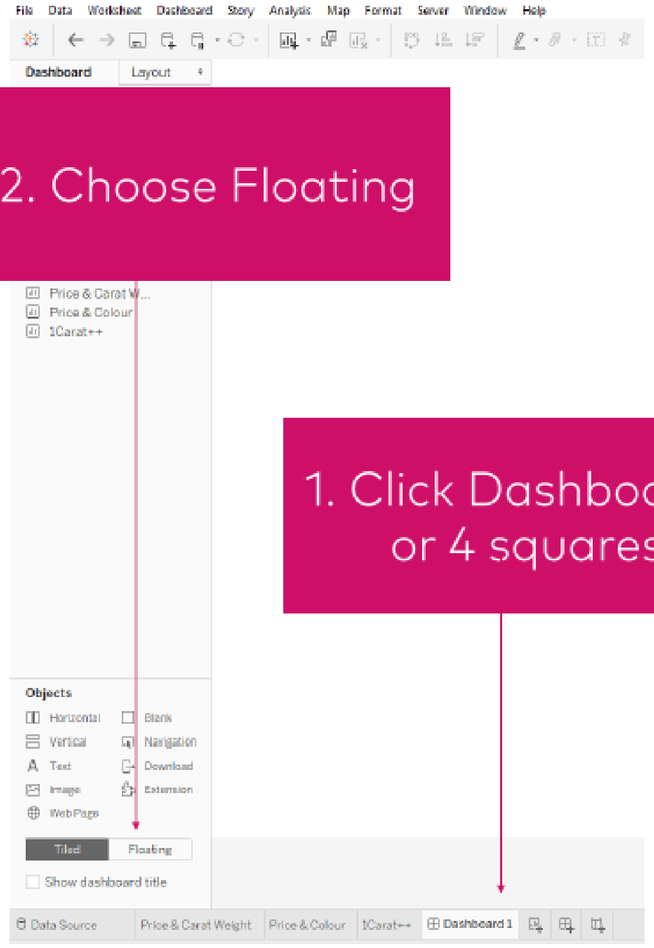
# LET'S DO ONE

1. Download the Tableau workbook "Module 5 Diamonds"
2. Open it and let's look through the 3 tabs I built
3. Our task is to put these graphs (and more) on a dashboard
4. In real life, I would draw this on paper first

# ADD A CHART TO THE DASHBOARD

**Instructions:**

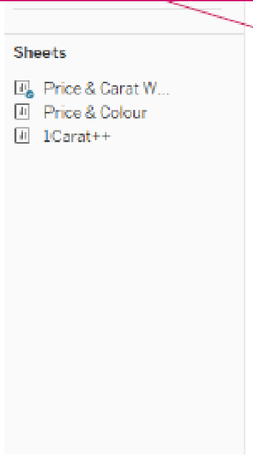
- On the bottom left, click on the tab called "Dashboard 1"
- On the left pane, under Size, change the dimension to be 1200x900
- On the bottom left, click Floating
- Drag Price & Carat weight to the canvas
- Do the same for the other tabs

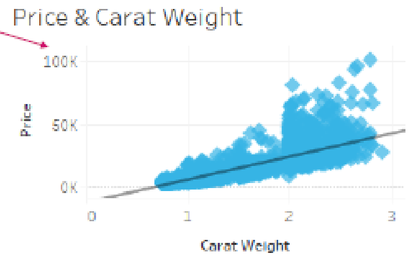



2. Choose Floating

1. Click Dashboard 1 or 4 squares

3. Drag "Price & Carat Weight" to anywhere on this dashboard canvas

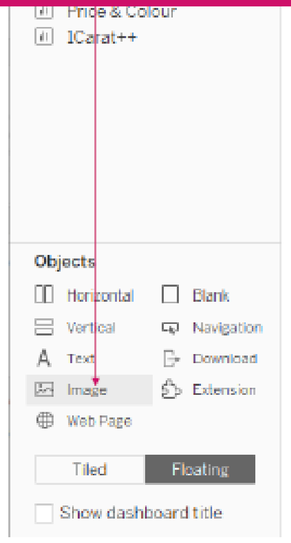


# SOME FUN THINGS

## 1. Add images

Click Image and navigate to the image you want to put on the dashboard

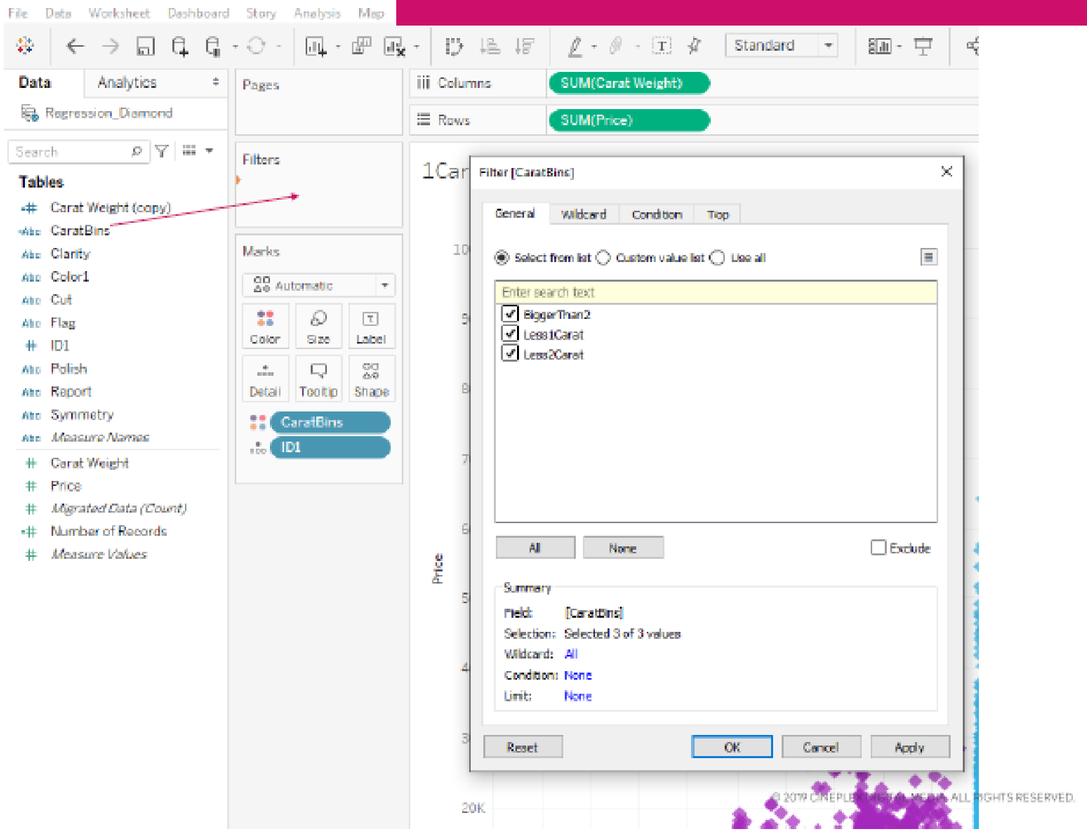



## 2. Add a title

Click Text and type in words (like a headline)

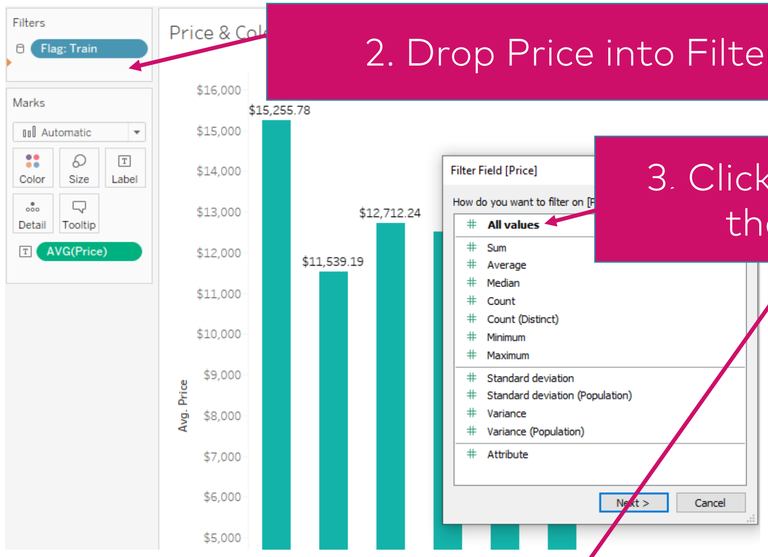

## 3. Add filters

Drag "CaratBins" to Filters



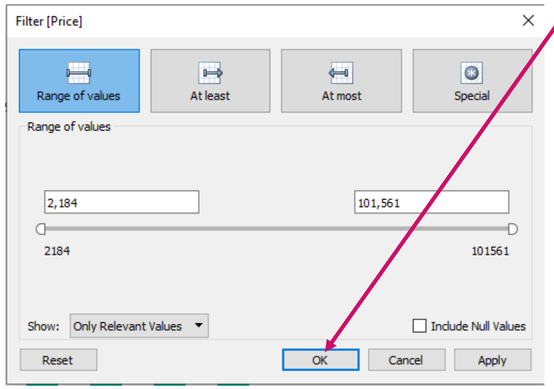
# UNDERSTANDING FILTERS

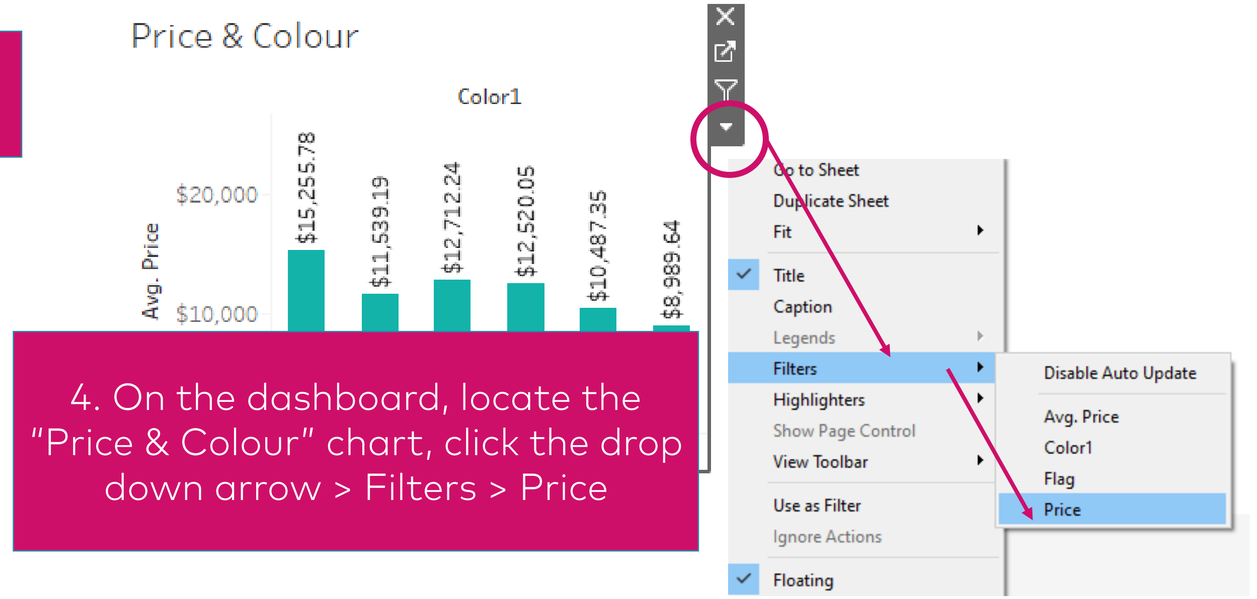
1. Go to the Price & Colour worksheet
2. Drag "price" to Filters
3. Click All values, hit OK
4. Go to the dashboard, click the drop down arrow > Filters > Price
5. The Price filter should appear on the dashboard
6. Apply to all charts, go to the drop down on the filter > Apply to worksheets > All Using this data source



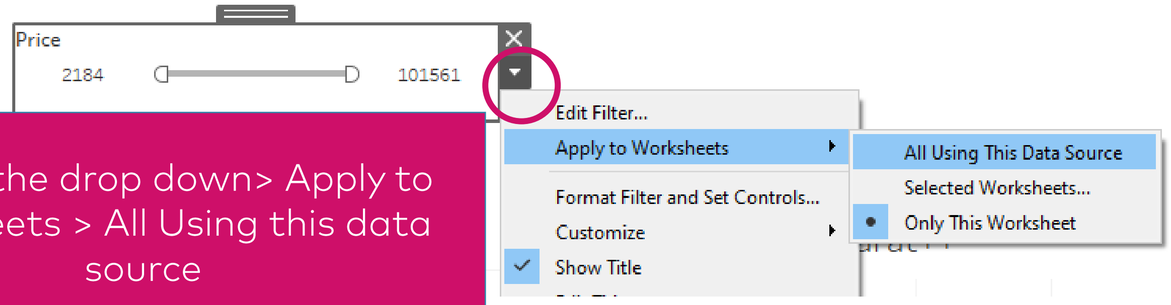

2. Drop Price into Filters

3. Click All Values then OK




4. On the dashboard, locate the "Price & Colour" chart, click the drop down arrow > Filters > Price




6. Click the drop down> Apply to worksheets > All Using this data source

# WRITING A CALCULATED FIELD

- Calculated fields allow you to create new data from data that already exists in your data source.

- When you create a calculated field, you are essentially creating a new field (or column) in your data source, the values or members of which are determined by a calculation that you control.

- This new calculated field is saved to your data source in Tableau, and can be used to create more robust visualizations. But don't worry: your original data remains untouched

https://help.tableau.com/current/pro/desktop/en-us/calculations_calculatedfields_create.htm

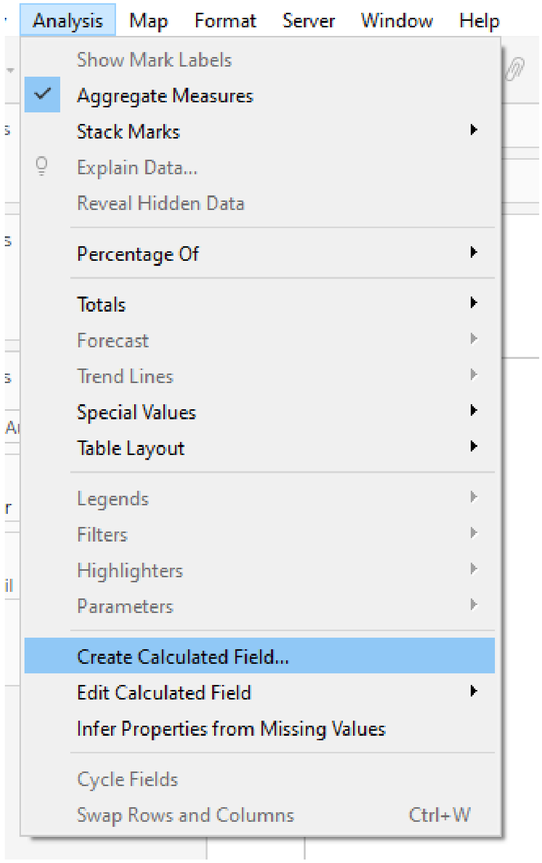
# CALCULATED FIELD: CARAT WEIGHT TO PRICE RATIO

**Objective:** Create a new data point called WP Ratio. This will measure the proportion of $ allocated to the weight of a diamond.

**Formula:** WP Ratio = Carat Weight/Price

**In Tableau:** Analysis > Create Calculated Field > Enter Name and formula

Hit OK




# ADD REFERENCE LINES

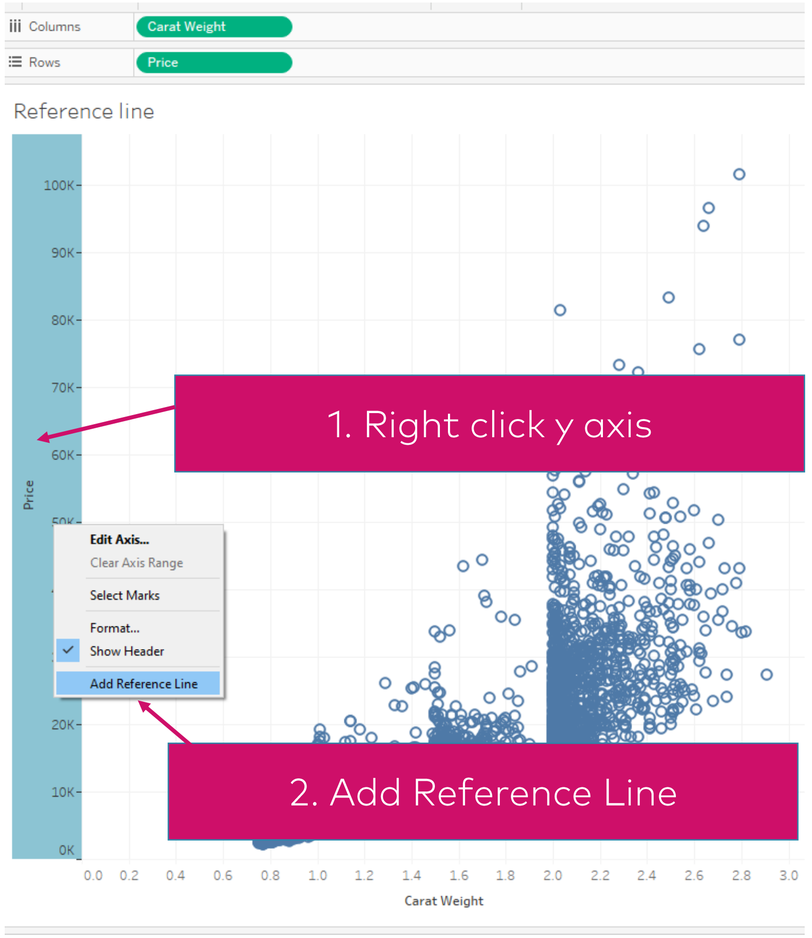
**Objective:** Add a reference line for the mean as a visual cue on the graph

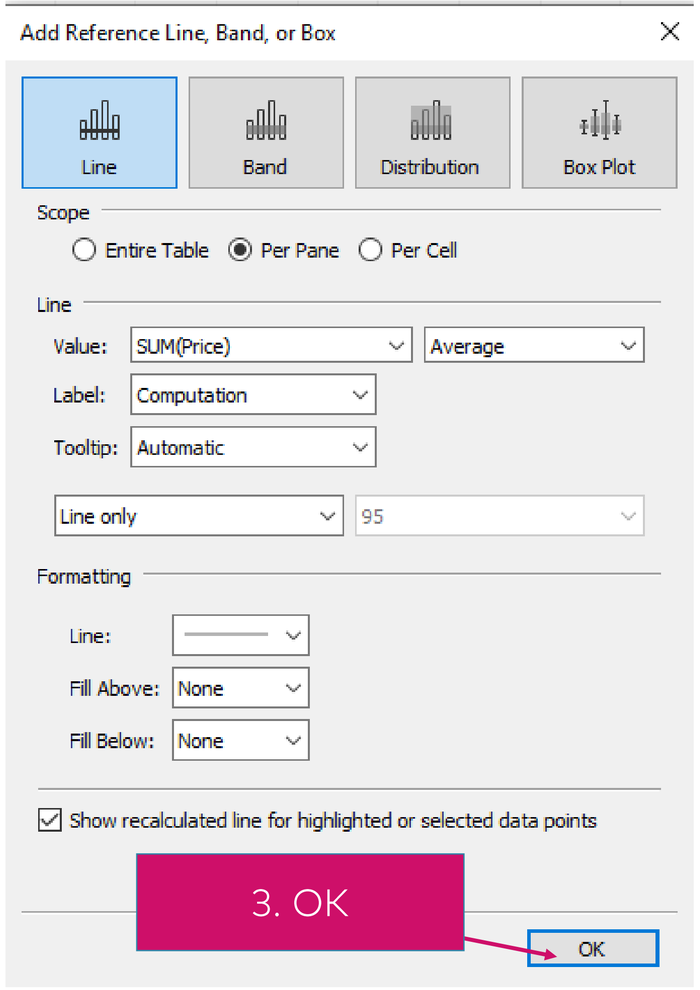
**In Tableau:** Plot Price on the y axis, Carat Weight on the x axis

Right click the y-axis

Click "Add reference line"

Click "OK"



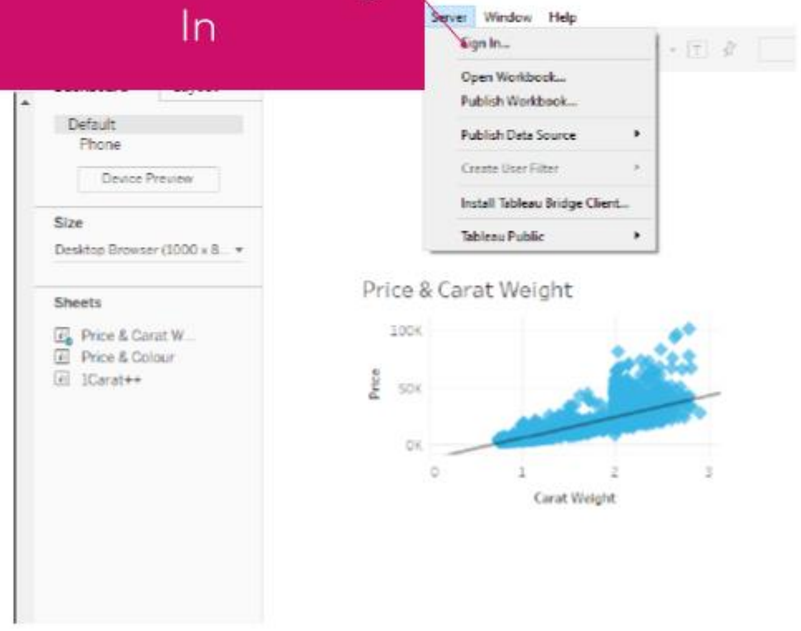



# POST YOUR DASHBORD

**Instructions:**

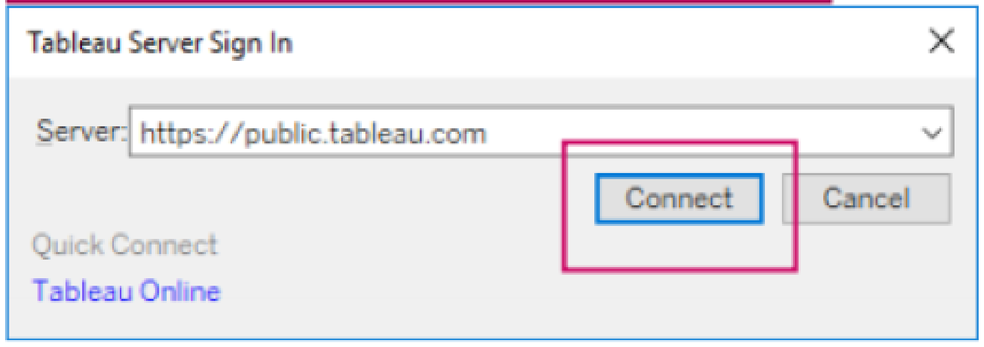
- Click Server > Sign In
- Type in https://public.tableau.com
- Click Connect and sign in (or create an account if you have not done so as yet)

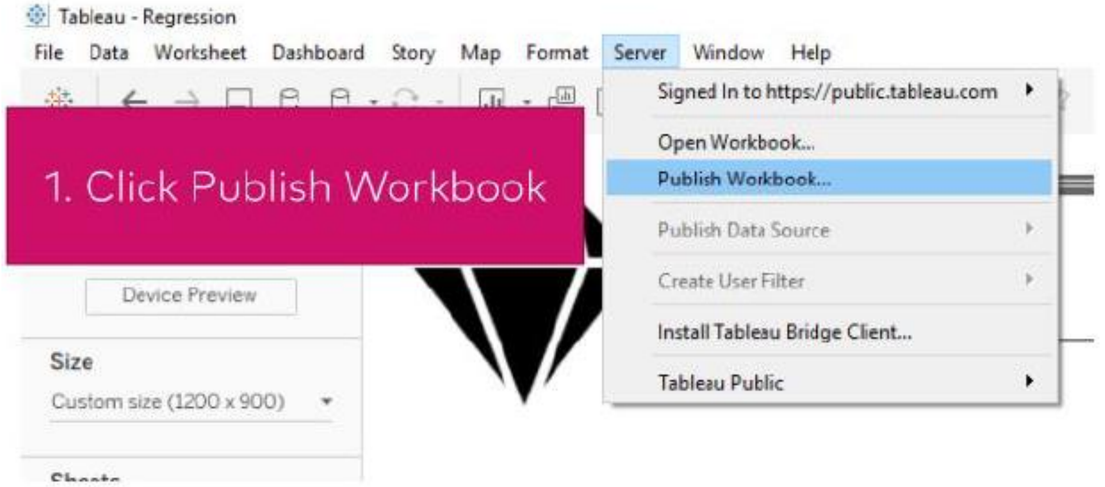
1. Click Server> Sign In

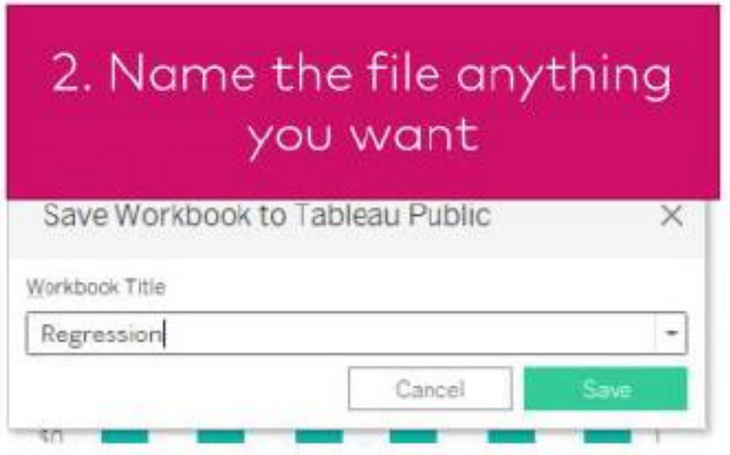



https://public.tableau.com

2. Type in https://public.tableau.com






**Instructions:**

- Click Server > Publish Workbook
- Name your dashboard something fun
- Hit Save

# IN CLASS EXERCISE

# YOUR TASK


**PRIZE:** $20 Gift certificate to Cineplex Store or Amazon

**YOUR TASK:**

1. Build and publish a dashboard on the diamonds dataset
2. The class will vote for a winner
3. You can work on your own or in a team ($20 is per entry so more people, less money for you)
4. I would suggest you do it individually for more ~~money~~ learning

**Entry due by Thursday 11th November 7pm**
We will publish it together in the beginning of class