**Quicksight analysis of Diamonds**

## Part 1 – Introduction to the data and materials

* The data we will use is a text file diamonds.txt which is a very simple file in structure (small number of columns and data types) but it will require some transformation and enrichment to make it more practical to analyse
* Diamonds attributes file explains the diamonds concepts which will be unfamiliar
* We have provided some images as supporting information – on clarity, colour and measurements: your challenge will be to consider when and what method to bring any supplementary information into your analysis

## Part 2 – transforming source data

* Try uploading as a new file - .txt is not supported - Supported file types include .csv, .tsv, .json, .clf, or .elf files.
* Text file formats are unstructured (ie there is nothing specific coded in to say when a new row or column begins. Quicksight requires that we first transform the file (parsing).
* Explain parsing –in general this means making something understandable. Parsing in this case means analysing a file, extracting essential information from it, into a structured/ usable format.
* We can parse the file with python, SQL, VBA, java methods, but for now lets use a SQL method because that is going to be useful for your next module
* Launch SQL server management studio (this is the client we will use to create and view databases
* Create a new database – sandbox
* In that database (right click) Tasks > import flat file
* Browse to location of file and set destination name as diamonds\_parsed
* Preview data
* Amend data types to be small and consistent ([brief explainer on sql data types](https://www.w3schools.com/sql/sql_datatypes.asp) – the main thing is to have something easy to use in QuickSight) all numericals to be decimal(8,2) 8 total characters, 2 dps, and all text to be varchar(20)
* Plus allow nulls on all columns (the file is too long to be certain there are no null values)
* Complete the wizard process and ensure this was successful
* Review the data using top 1000 rows
* OPTIONAL In SQL we will add a row id to the table which will be often useful for analysis and because tables held in SQL normally have a primary key or unique value to identify each row.

**ALTER TABLE diamonds\_parsed**

**ADD rowid INT IDENTITY(1,1) NOT NULL**

* Now we will export the data into the new usable csv format supported by QuickSight
* Query the table in SQL server management studio to show the full data in the table

**Select \* from diamonds\_parsed**

* Select anywhere in the results and right click > export results to csv
* Save that file in your diamonds folder as a csv format
* Open the file in excel to confirm the file looks as expected.
* Filter the columns to see if we have any bad or missing data (hint : xyz contains 0s which indicates some data may have been dropped from the source during the parsing process)
* Make notes on these and any other findings. Close the file without saving.
* Connect to this data set in QuickSight (import csv to SPICE and visualise)
* Create a quick autograph of clarity (raw data)

A graph of blue bars

Description automatically generated

## Part 3 – enriching data methods

* Option 1 - prepare a lookup table – example : in excel, lets prepare a table which links clarity to definition and provides an order for them, which we could import as an additional source for the data set.

A screenshot of a table

Description automatically generated

* Edit data set
* Add data > upload file > browse to file location
* Inner join, on clarity > apply
* Publish and visualise to return to your analysis
* Replace clarity in the original visual with clarity label
* Sort by order

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* Option 2 – add calculations inside QuickSight – example: in QuickSight we will add a calculated column which assigns colour labels to codes

ifelse(color='D' OR color='E' OR color ='F', 'colourless', 'near-colourless')

* Option 3 - we could also consider editing the source (but note this data has 54k rows so is going to be hard to manage in excel)

## Part 3 – set up for the tasks.

* Learners to open task file and run through intro paragraph. Underline the aim is not to complete everything but pay close attention to the details and the design.
* Tell them how much time they have – this task starts day 2 pm and finishes with a show and tell/ Q&A on day 3 am
* Encourage them to consider designing a dashboard to solve groups of tasks in one layout
* Remind them that they can work iteratively – consider the tasks to be the requirements and discover, through iterations, better ways to provide the information through QuickSight