To,

Sprocket Central Pty Ltd,

Subject: Data Quality Review

Hi there,

I want to state that the datasets provided by you is so raw and have some quality issues. I have been found that some of the field has incorrect values, issues with the completeness, inconsistency of data, irrelevancy and issue with the valid data. So, we need to work out and refine the data, analyse and process the data to drive value driven results in the business.

I have attached below some of the quality issues in the dataset provided and refine those data to qualify it for further process.

Kind regards

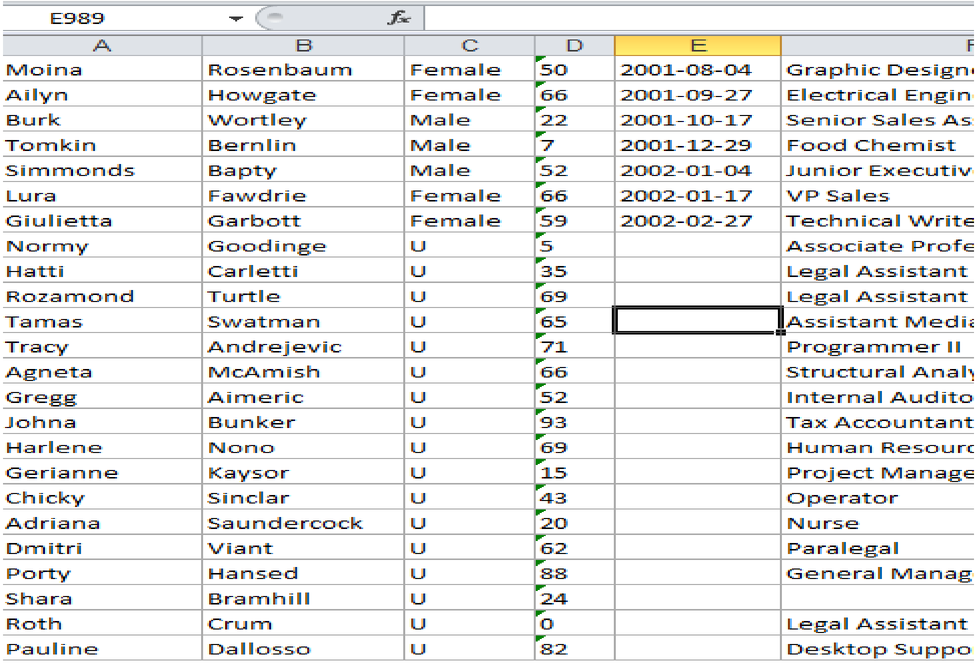
Binod Kshetry

Data Analytics & Consulting Intern (Virtual)

KPMG

Various columns such as job title, date of birth (DOB), brand of purchase have empty value in certain records. **Incompleteness** of data is seen as some of the value in certain fields are missing.

Mitigation: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

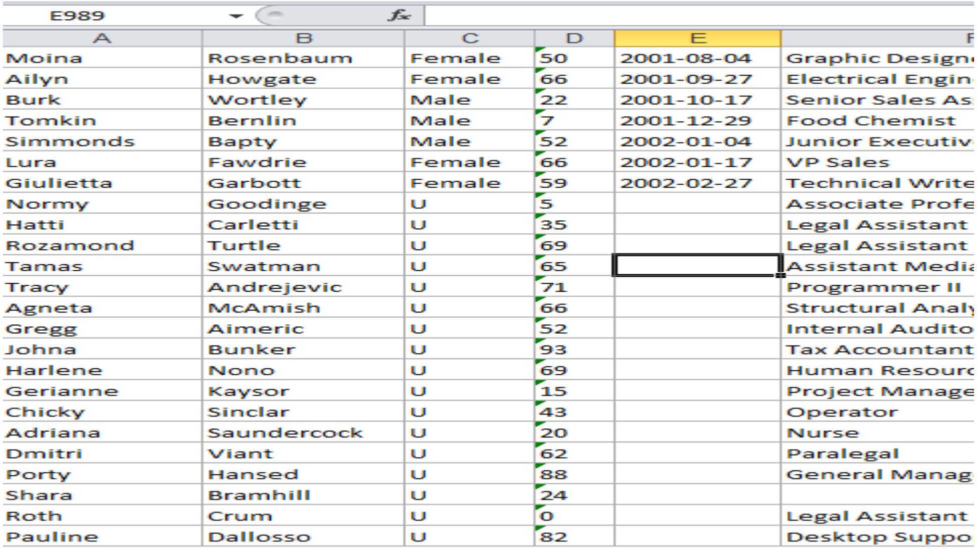


E989 *fx*

A	B	C	D	E	F
Moina	Rosenbaum	Female	50	2001-08-04	Graphic Design
Ailyn	Howgate	Female	66	2001-09-27	Electrical Engin
Burk	Wortley	Male	22	2001-10-17	Senior Sales As
Tomkin	Bernlin	Male	7	2001-12-29	Food Chemist
Simmonds	Bapty	Male	52	2002-01-04	Junior Executiv
Lura	Fawdrie	Female	66	2002-01-17	VP Sales
Giulietta	Garbott	Female	59	2002-02-27	Technical Write
Normy	Goodinge	U	5		Associate Profe
Hatti	Carletti	U	35		Legal Assistant
Rozamond	Turtle	U	69		Legal Assistant
Tamas	Swatman	U	65		Assistant Media
Tracy	Andrejevic	U	71		Programmer II
Agneta	McAmish	U	66		Structural Analy
Gregg	Aimeric	U	52		Internal Audito
Johna	Bunker	U	93		Tax Accountant
Harlene	Nono	U	69		Human Resourc
Gerianne	Kaysor	U	15		Project Manage
Chicky	Sinclar	U	43		Operator
Adriana	Saundercock	U	20		Nurse
Dmitri	Viant	U	62		Paralegal
Porty	Hansed	U	88		General Manag
Shara	Bramhill	U	24		
Roth	Crum	U	0		Legal Assistant
Pauline	Dallosso	U	82		Desktop Suppo

Also, Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)

Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses. Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. Additionally,



A	B	C	D	E	F
Moina	Rosenbaum	Female	50	2001-08-04	Graphic Design
Ailyn	Howgate	Female	66	2001-09-27	Electrical Engin
Burk	Wortley	Male	22	2001-10-17	Senior Sales As
Tomkin	Bernlin	Male	7	2001-12-29	Food Chemist
Simmonds	Bapty	Male	52	2002-01-04	Junior Executiv
Lura	Fawdrie	Female	66	2002-01-17	VP Sales
Giulietta	Garbott	Female	59	2002-02-27	Technical Write
Normy	Goodinge	U	5		Associate Profe
Hatti	Carletti	U	35		Legal Assistant
Rozamond	Turtle	U	69		Legal Assistant
Tamas	Swatman	U	65		Assistant Medi
Tracy	Andrejevic	U	71		Programmer II
Agneta	McAmish	U	66		Structural Analy
Gregg	Aimeric	U	52		Internal Audito
Johna	Bunker	U	93		Tax Accountant
Harlene	Nono	U	69		Human Resourc
Gerianne	Kaysor	U	15		Project Manage
Chicky	Sinclar	U	43		Operator
Adriana	Saundercock	U	20		Nurse
Dmitri	Viant	U	62		Paralegal
Porty	Hansed	U	88		General Manag
Shara	Bramhill	U	24		
Roth	Crum	U	0		Legal Assistant
Pauline	Dallosso	U	82		Desktop Suppo

So, we need to work on data cleaning, standardization and transformation process for the purpose of model analysis. Assumptions will be made based on the queries raised. After completion, your data will be analysed further based on assumptions made to come up with meaningful insights for better decision making.