Project 2.1: Data Cleanup

Faça uma cópia deste documento. Complete cada seção. Quando estiver pronto, salve seu arquivo como um documento PDF e envie-o aqui: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Passo 1: Entendimento do Negócio e dos Dados

Sugerir qual a cidade ideal para abrir um novo petshop da Pawdacity.

Decisões Chave:

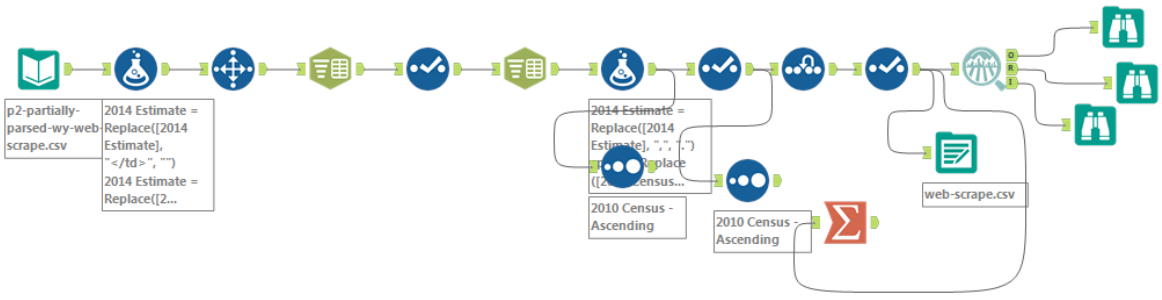
Responda estas perguntas

1. Que decisões devem ser tomadas?
   Através da verificação de determinados fatores (densidade demográfica, população maior de 18 anos, dados históricos sobre vendas dos petshop's da loja), identificar qual a melhor cidade para abrir um novo negócio.

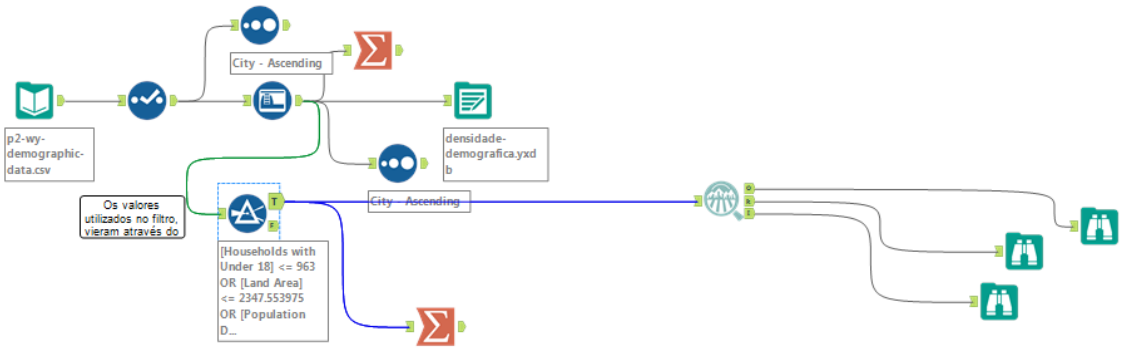
2. Que dados são necessários para subsidiar essas decisões?
   No exercício são passados para nós 4 conjuntos de dados. Para a primeira parte de data cleaning são necessários apenas 3 destes conjuntos: o que mostra um web scrapping com informações sobre as cidades (censo populacional), um outro que mostra a densidade populacional e por fim um histórico de vendas.

Passo 2: Construindo o Conjunto de Treinamento

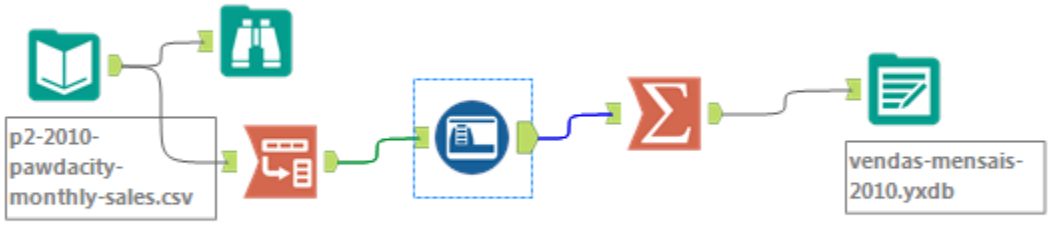
A primeira parte consistia-se em realizar a limpeza dos dados, primeiramente fiz com os dos dados de web scrap. Abaixo segue imagem do fluxo realizado. Como pode ser visto, ao final do processo foi gerado o arquivo "web-scrape.csv"



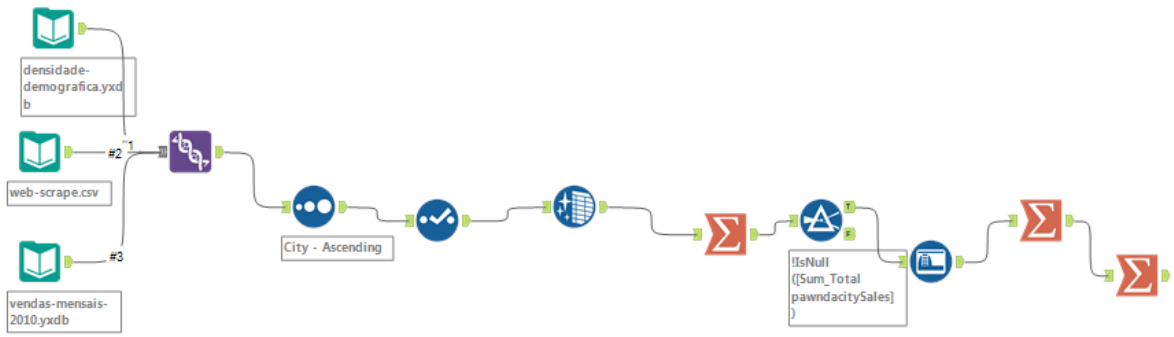
Após isso foi realizada a limpeza dos dados da densidade demográfica, conforme o gráfico abaixo:



A primeira vista, como pode ser analisado na imagem, eu havia feito o tratamento para outliers utilizando o método IQR, porém ao fazer isso a quantidade de dados diminuiu bastante.
Por fim, o último modelo apresentava todas as vendas do ano de 2010, mês a mês, então utilizei a ferramenta de transposição para colocar todas as colunas em apenas uma e somar o valor, ficando assim mais fácil de ser analisada.



Após gerar os 3 documentos, importei eles, realizei um union, limpei os arquivos, retirando espaços, fiz um group by pela cidade e somei o valor das demais colunas.



Como uma das colunas importantes é a Total Sales, e existiam apenas 11 registros dela, fiz um filtro para pegar os campos onde a mesma não era nula.

Column	Sum	Average
Census Population	*213,862*	*19.442*
Total Pawdacity Sales	*3,773,304*	*343027.63*
Households with Under 18	*34,064*	*3096.72*
Land Area	*33,071*	*3006.48*
Population Density	*63*	*5.70*
Total Families	*62,653*	*5695.70*

Passo 3: Tratando os Outliers

Existem cidades que são outliers no conjunto de treinamento? Qual outlier você escolheu para remover ou imputar? Como esse conjunto de dados é um conjunto de dados pequeno (11 cidades), **você deve apenas remover ou imputar um outlier**. Explique o seu raciocínio.

Existem, para remover as cidades que são outliers utilizei o método IQR. A condição presente no filtro ficou da seguinte maneira:

[Households with Under 18] <= 963 OR [Land Area] <= 2347.553975 OR [Population Density] <= 2.315 OR [Total Families] <= 2024.385

O cálculo utilizado nos retorna 5 registros que não se adequam a essa equação (caso queira conferir os cálculos utilizados para essa equação, o mesmo está presente em anexo com o nome: Densidade Populacional.xlsx)
Foram as seguintes cidades:

City	County	Land Area	Households with Under 18	Population Density	Total Families
Laramie	Albany	2513.745235	2075	5.19	4668.93
Gillette	Campbell	2748.8529	4052	5.8	7189.43
Riverton	Fremont	4796.859815	2680	2.34	5556.49
Casper	Natrona	3894.3091	7788	11.16	8756.32
Rock Springs	Sweetwater	6620.201916	4022	2.78	7572.18

Ao analisar os outliers, preferi remover a cidade Laramie, pois ao comparar com os valores presentes no outro conjunto de dados (Pawdacity monthy sales) ele é o único dos presentes que não possui dados lá.

Antes de enviar

Por favor, verifique suas respostas contra os requisitos do projeto ditados por esta rubrica usada pelos revisores para classificar seu projeto.