Catalogue de données pour le projet de Pipeline d'analyse de données de vente de voitures

Description des données :

Les données proviennent des deux sources suivantes :

1. DVM-CAR

: Cet ensemble de données accessible au public vise à faciliter la recherche et les applications commerciales dans l'industrie automobile, telles que la conception de l'apparence des voitures, l'analyse des consommateurs et la modélisation des ventes. **Fichiers :**

- Tableau de base : attributs des voitures tels que le nom du modèle, l'identifiant du modèle et le nom de la marque.
- Tableau des ventes : données de ventes de voitures sur dix ans au Royaume-Uni.
- Tableau des prix : prix des voitures neuves au fil des ans pour le niveau d'entrée.
- Tableau de garniture : attributs de garniture comme le prix de vente (niveau de garniture), le type de moteur et la taille du moteur.
- Tableau des annonces : plus de 0,25 million d'annonces de voitures d'occasion.
- Tableau d'images : attributs des images de voitures tels que la couleur et le point de vue.

2. Marketcheck API

Balise globale du site (gtag.js) - Google Analytics

Marketcheck est une source principale de divers ensembles de données web automobiles fournis via nos flux de données et API.

L'API Marketcheck offre un accès facile aux annonces de véhicules en ligne aux États-Unis et au Canada, ainsi qu'à des sources de données connectées provenant de plusieurs domaines verticaux.

La plateforme API Marketcheck gère plus de 60 millions d'appels API par mois provenant de plus de 130 abonnés et est devenue une norme dans le secteur automobile en Amérique du Nord.

APIs : Voitures, Motocyclettes, Véhicules récréatifs (RV), Équipements lourds.

Données utilisées dans le projet :

DVM-CAR

Nous utiliserons le tableau des ventes, qui est une jointure des différentes tables et comporte au total 268255 lignes et 24 colonnes, à savoir : Maker, Genmodel, GenmodelID, AdvID, Advyear, Advmonth, Color, Regyear,

Bodytype, RunnedMiles, Enginsize, Gearbox, Fueltype, Price, Enginepower, AnnualTax, Wheelbase,

Height, Width, Length, Averagempg, Topspeed, Seatnum, Doornum

Marketcheck API

Nous utiliserons l'API Cars avec le point "Inventory search". **market**:

- **id**
- **vin**
- **heading**
- **price**
- **miles**
- **msrp**
- **data_source**
- **vdp_url**
- **carfax_1_owner**
- **carfax_clean_title**
- **exterior_color**
- **interior_color**
- **dom**
- **dom_180**
- **dom_active**
- **seller_type**
- **inventory_type**
- **stock_no**
- **last_seen_at**
- **last_seen_at_date**
- **scraped_at**
- **scraped_at_date**
- **first_seen_at**
- **first_seen_at_date**
- **ref_price**
- **ref_miles**
- **source**
- **dealer**:
  - **id**
  - **website**
  - **name**
  - **dealer_type**
  - **street**
  - **city**
  - **state**

  - **country**
  - **latitude**
  - **longitude**
  - **zip**
  - **phone**
- **build**:
  - **year**
  - **make**
  - **model**
  - **trim**
  - **short_trim**
  - **vehicle_type**
  - **transmission**
  - **drivetrain**
  - **fuel_type**
  - **engine**
  - **engine_size**
  - **engine_block**
  - **doors**
  - **cylinders**
  - **made_in**
  - **steering_type**
  - **antibrake_sys**
  - **tank_size**
  - **overall_height**
  - **overall_length**
  - **overall_width**
  - **std_seating**
  - **opt_seating**
  - **trim_r**

Données finales :

En se basant sur les objectifs de notre projet, nous avons choisi de travailler avec ces données : **Maker Model Adv_year Adv_month Color Reg_year Bodytype Miles Engine_size Gearbox Fuel_type Price Height Width Length Average_mpg Seatings Doors Stat country ID**

Traitement des données

Les données seront colectées depuis les differentes sources mentionnés ici, seront mergé puis sauvegardé dans une zone de transit qui sera dans HDFS. Nouos appliquerons des transformations métiers à partir de HDFS, ferons la modélisation et les sauvegarderons dans SQL Server en tant qu'entrepôt de données. Puis nous les utiliserons dans PowerBI pour faire les analyses.

Date de publication : 05/01/2024
Signé : Sefdine Nassuf