Projet 5 DataAnalyst OCR

Sommaire

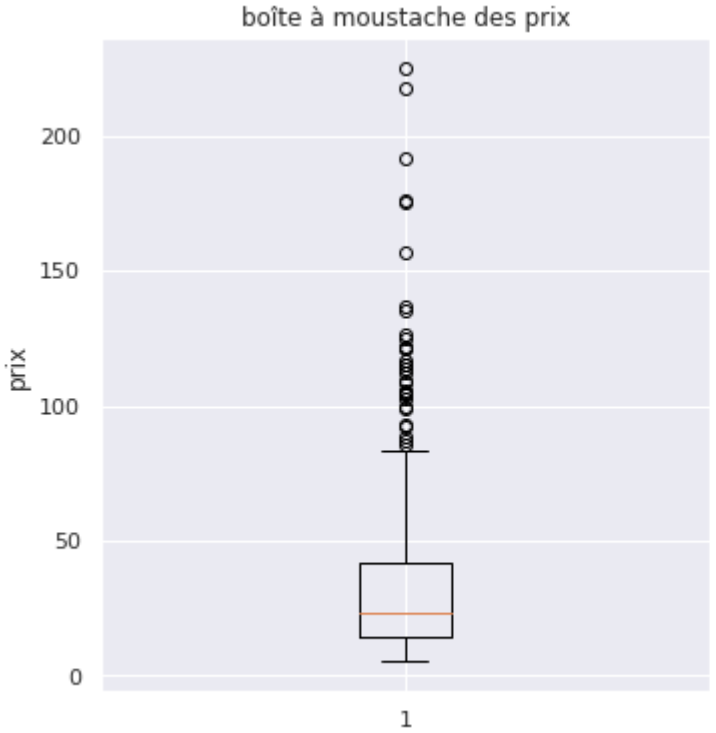
- Chiffre d’affaires par produit et total en ligne
- Boxplot et Histogramme des prix
- Courbe de Lorentz
- Liste des prix aberrants
- Boxplot catégorie
- Listes des prix aberrants par catégorie (exemple “vin”)
- Boxplot des ventes aberrantes
- Liste des ventes aberrantes
- Codage : calcul de normalisation
- Codage : matplotlib
- Codage : création de catégorie

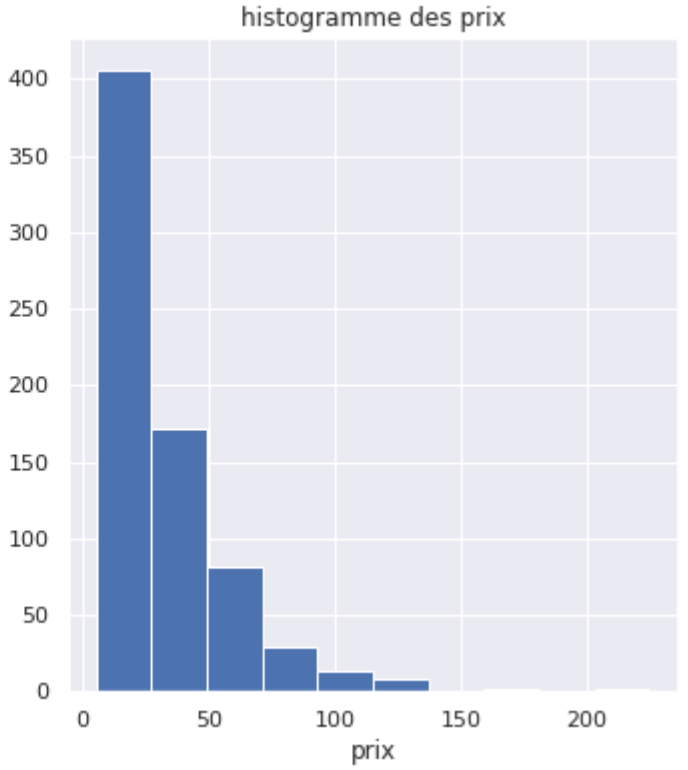
Chiffre d’affaires par produit

	product_id	price	onsale_web	stock_quantity	sku	total_sales	post_name	CA
286	4334	49.0	1	0	7818	96.0	champagne-gosset-grand-blanc-de-blanc	4704.0
162	4144	49.0	1	11	1662	87.0	champagne-gosset-grand-rose	4263.0
310	4402	176.0	1	8	3510	13.0	cognac-frapin-vip-xo	2288.0
161	4142	53.0	1	8	11641	30.0	champagne-gosset-grand-millesime-2006	1590.0
160	4141	39.0	1	1	304	40.0	gosset-champagne-grande-reserve	1560.0
293	4355	126.5	1	2	12589	11.0	champagne-egly-ouriet-grand-cru-brut-blanc-de-...	1391.5

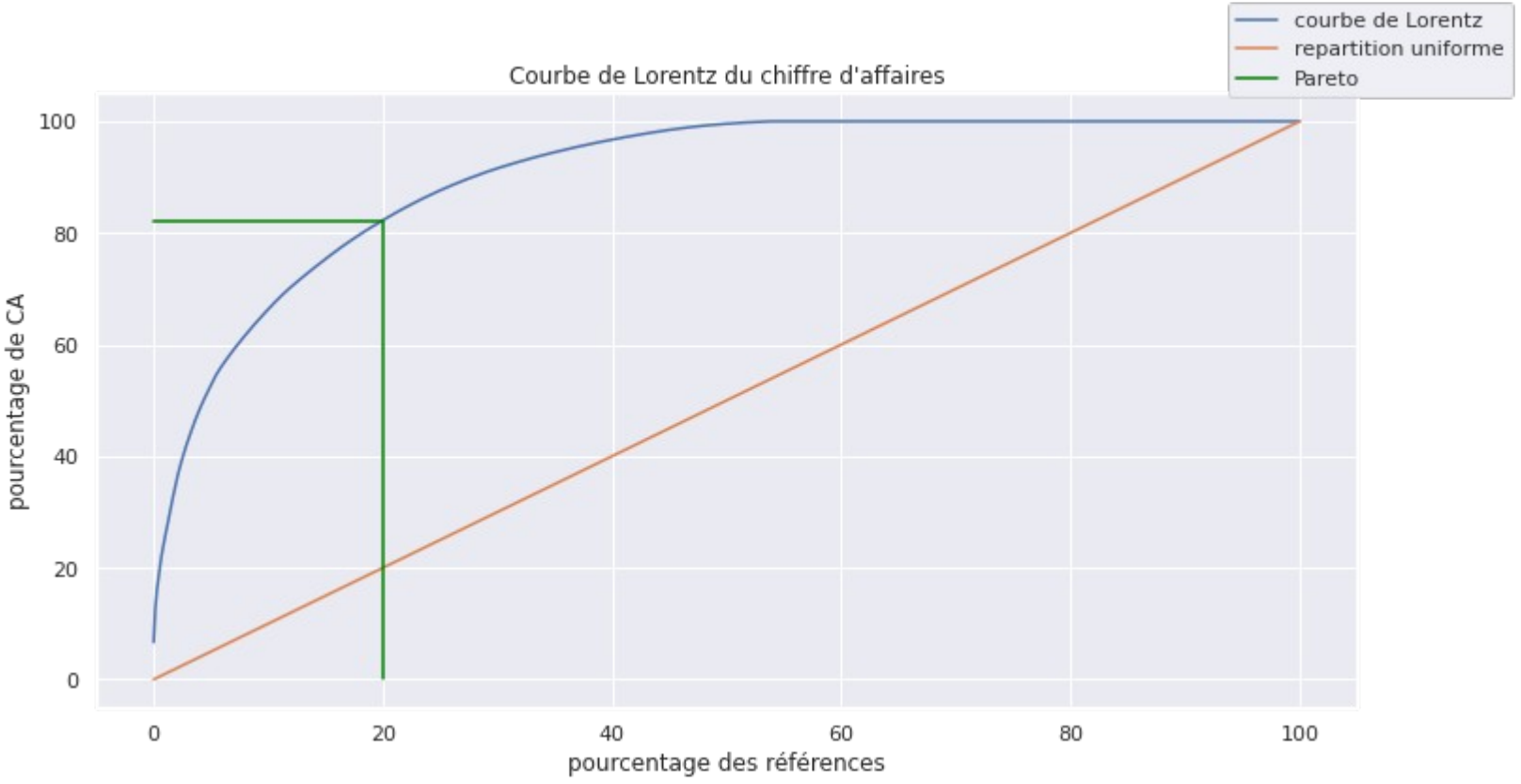
```
70568.6 euros de chiffre d'affaires web
```

Boxplot et Histogramme des prix





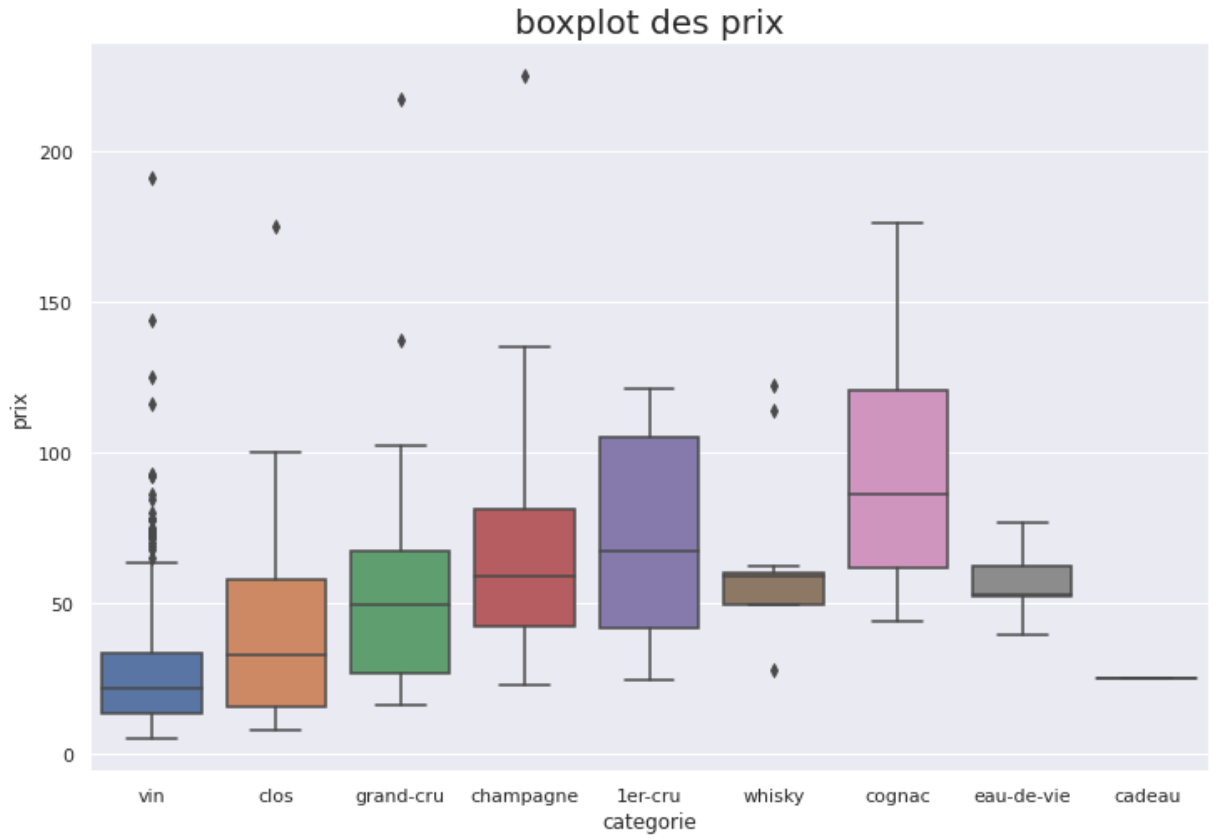
Courbe de Lorentz



Liste des prix aberrants

	product_id	price	CA	post_name
291	4352	225.0	1125.0	champagne-egly-ouriet-grand-cru-millesime-2008
525	5001	217.5	0.0	david-duband-charmes-chambertin-grand-cru-2014
692	5892	191.3	573.9	coteaux-champenois-egly-ouriet-ambonnay-rouge-...
310	4402	176.0	2288.0	cognac-frapin-vip-xo
657	5767	175.0	0.0	camille-giroud-clos-de-vougeot-2016
313	4406	157.0	0.0	cognac-frapin-chateau-de-fontpinot-1989-20-ans
30	4594	144.0	NaN	post_name_vide
478	4904	137.0	685.0	domaine-des-croix-corton-charlemagne-grand-cru...
752	6126	135.0	270.0	champagne-gosset-celebris-vintage-2007
293	4355	126.5	1391.5	champagne-egly-ouriet-grand-cru-brut-blanc-de-...

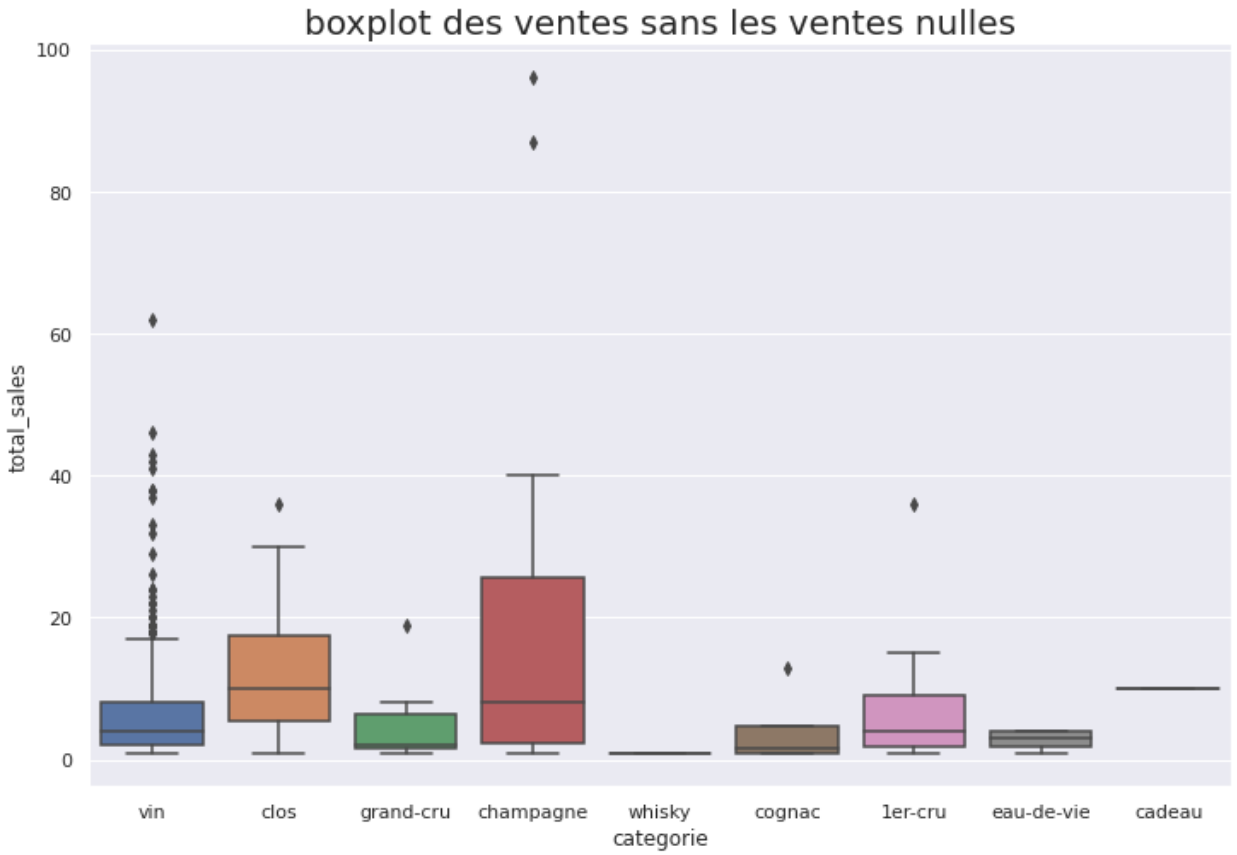
Boxplot par catégorie



Listes des prix aberrants par catégorie (exemple “vin”)

	product_id	price	total_sales	stock_quantity	post_name
692	5892	191.3	3.0	10	coteaux-champenois-egly-ouriet-ambonnay-rouge-...
30	4594	144.0	NaN	0	post_name_vide
615	5612	124.8	0.0	12	domaine-weinbach-gewurztraminer-gc-furstentum-...
758	6202	116.4	0.0	14	domaine-clerget-echezeaux-en-orveaux-2015
707	5916	93.0	0.0	3	wemyss-malts-single-cask-chocolate-moka-cake
55	6324	92.0	NaN	18	post_name_vide
605	5565	92.0	0.0	0	tempier-bandol-cabassaou-2017
19	4055	86.1	NaN	0	post_name_vide
47	5070	84.7	NaN	0	post_name_vide
10	4046	80.0	6.0	0	pierre-gaillard-cote-rotie-rose-pourpre-2017
523	4996	78.0	0.0	33	domaine-peyre-rose-marlene-n3-2008
522	4995	78.0	0.0	0	domaine-peyre-rose-oro-2002
521	4994	78.0	0.0	7	domaine-peyre-rose-syrah-leone-2008

Boxplot des ventes



Listes des ventes aberrantes

	product_id	price	total_sales	stock_quantity	post_name
692	5892	191.3	3.0	10	coteaux-champenois-egly-ouriet-ambonnay-rouge-...
30	4594	144.0	NaN	0	post_name_vide
615	5612	124.8	0.0	12	domaine-weinbach-gewurztraminer-gc-furstentum-...
758	6202	116.4	0.0	14	domaine-clerget-echezeaux-en-orveaux-2015
707	5916	93.0	0.0	3	wemyss-malts-single-cask-chocolate-moka-cake
55	6324	92.0	NaN	18	post_name_vide
605	5565	92.0	0.0	0	tempier-bandol-cabassaou-2017
19	4055	86.1	NaN	0	post_name_vide
47	5070	84.7	NaN	0	post_name_vide
10	4046	80.0	6.0	0	pierre-gaillard-cote-rotie-rose-pourpre-2017
523	4996	78.0	0.0	33	domaine-peyre-rose-marlene-n3-2008
522	4995	78.0	0.0	0	domaine-peyre-rose-oro-2002
521	4994	78.0	0.0	7	domaine-peyre-rose-syrah-leone-2008
126	4073	77.8	0.0	11	chateau-de-vaudieu-chateauneuf-du-pape-lavenue...

Codage : calcul de normalisation

```
#dataframe ou tableaux avec valeur abérrante en sortie

data['score'] = (data['price']-data['price'].mean())/data['price'].std()
display(data[abs(data['score'])>ecart_significatif][['product_id','price','CA','post_name']]
        .sort_values(by='price',ascending=False).head(15))
```

```
#même chose que précédemment avec la bibliothèque sklearn
sc = StandardScaler() #création d'un objet StandartScaler
data['score2'] = sc.fit_transform(data[['price']])
display(data[abs(data['score2'])>ecart_significatif][['product_id','price','CA','post_name']]
        .sort_values(by='price',ascending=False).head(15))
```

Codage : matplotlib

```
#theme visuel du graphe
sns.set_theme(style='darkgrid')

#creation objet pyplot figure et axe (en mode OOP)
fig , ax = plt.subplots(1,2)

#agrandissement taille figure
fig.set_size_inches(12, 6)

#creation d'une boite à moustache
ax[0].boxplot(data['price'])
ax[0].set_title('boite a moustache des prix')
ax[0].set_ylabel('prix')

# creation d'un histogramme
ax[1].hist(data['price'])
ax[1].set_title('histogramme des prix')
ax[1].set_xlabel('prix')

# boucle pour modifier les differents graphes
for a in ax:
    a.yaxis.grid(True)

plt.show() #pas utile sur jupyter mais j'aime bien
```

Codage : création de catégorie

```
#creation d'une colonne data avec valeurs par defaut 'vin'
data['categorie'] = 'vin'

#creation d'un dictionnaire avec les mot recherchés et leurs categories
dic_categorie = {'grand-cru':'grand-cru',
                 'champagne':'champagne',
                 '1er-cru':'1er-cru',
                 '1cru':'1er-cru',
                 '1ercru':'1er-cru',
                 'clos':'clos',
                 'cognac':'cognac',
                 'whisky':'whisky',
                 'eau-de-vie':'eau-de-vie',
                 'cadeau':'cadeau'
                }

#changer la categorie en fonction d'un mot contenu dans la chaine de caractere
for cle,cat in dic_categorie.items():
    perso.categoriser(data,'post_name',cle,cat,regex=False)

#affichage des categories trouvées
#for cat in data['categorie'].unique():
#    display(data[data['categorie'] == cat])

#plus besoin du dictionnaire: effacement en memoire
del dic_categorie
```