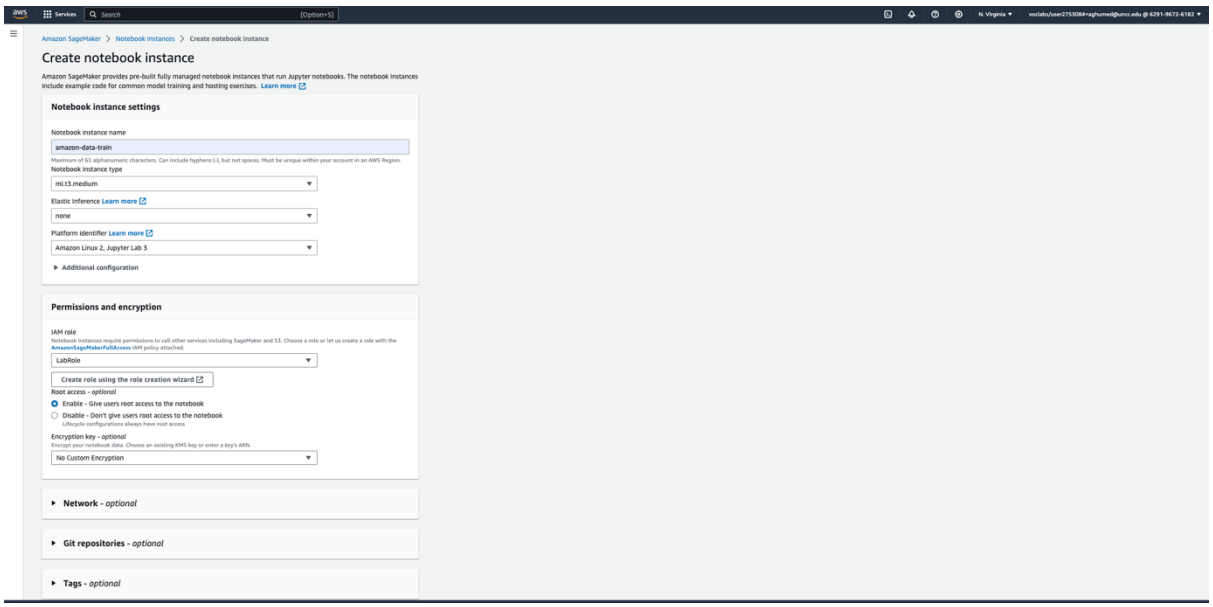
Project Deliverable 3

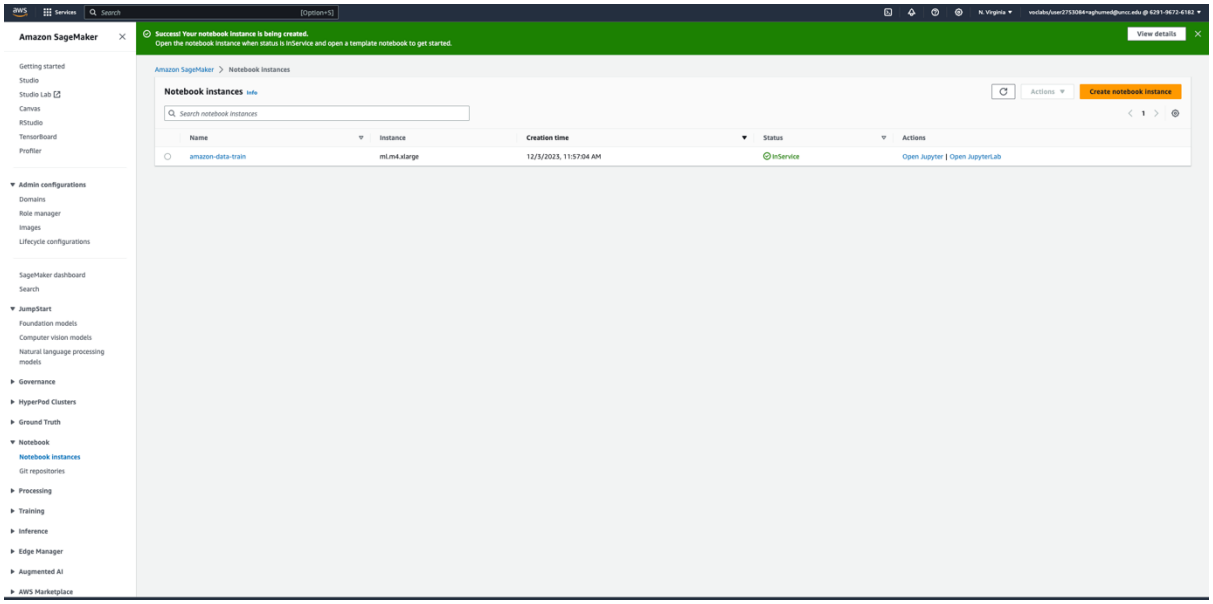
Team-6

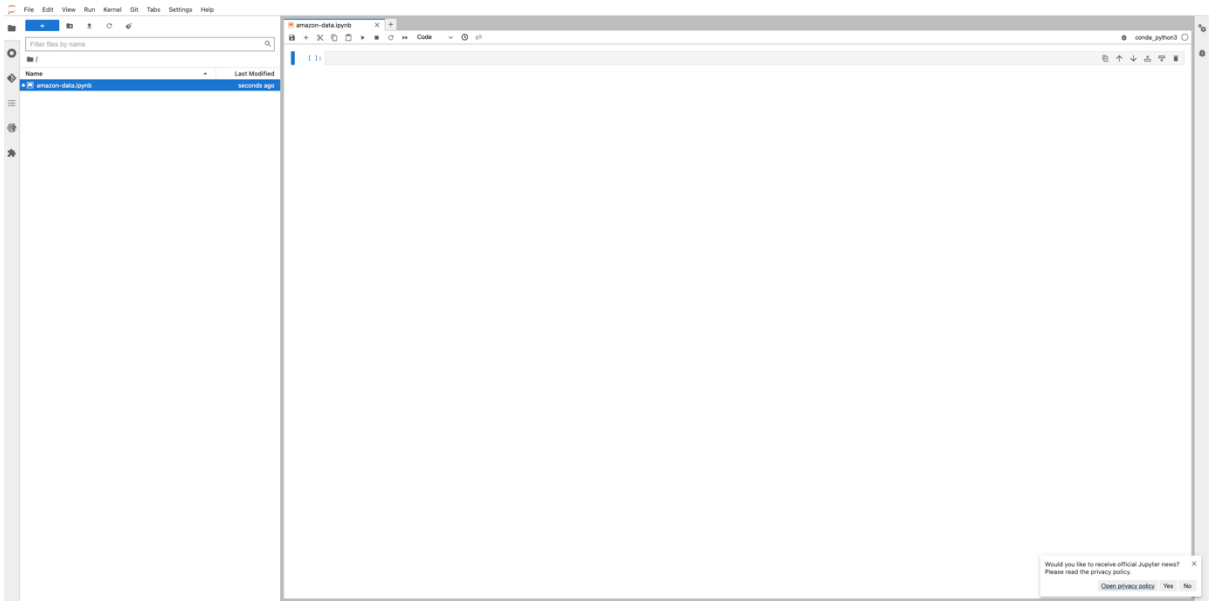
Team Members:

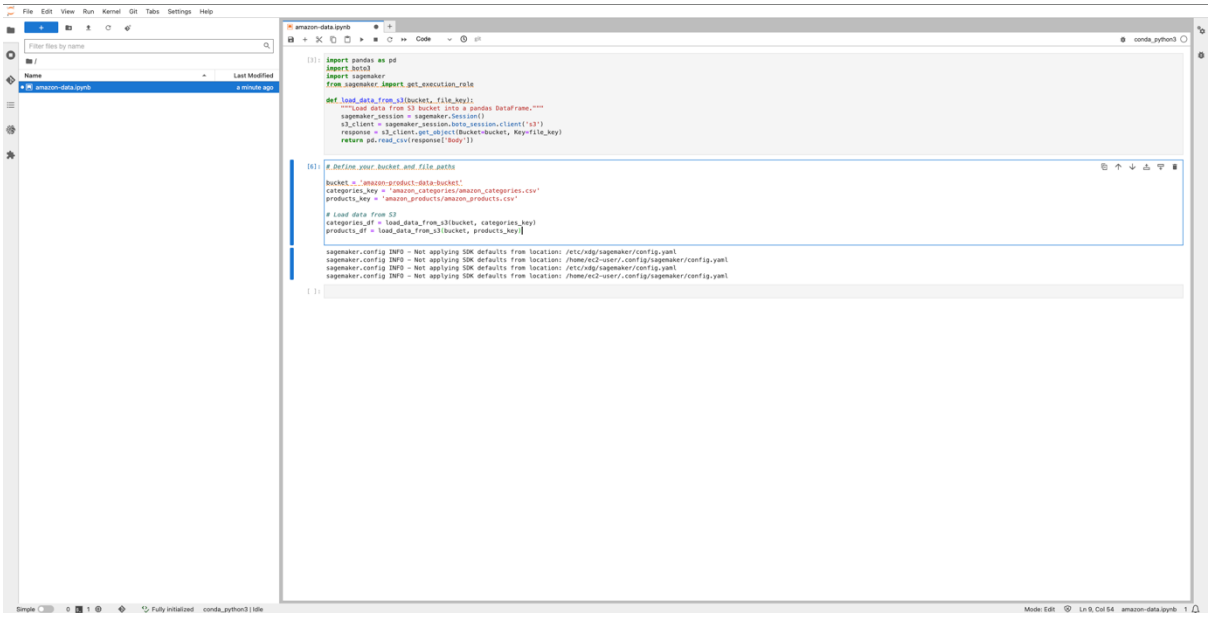
1. Aniket Ghumed
2. Anantnaval Gaikward
3. Rimjhim Jain
4. Saumit Chinchkandi
5. Sudhanshu Dalvi

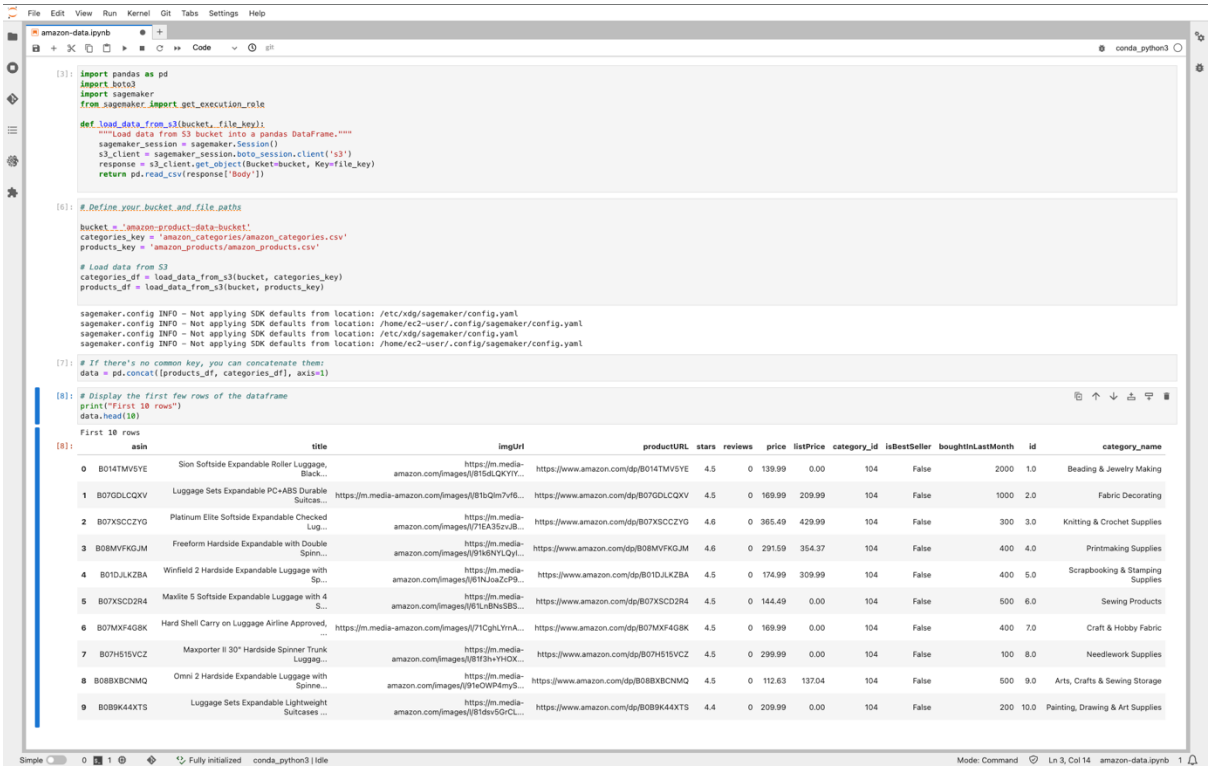
PHASE 3 Modeling, Evaluation, Tuning



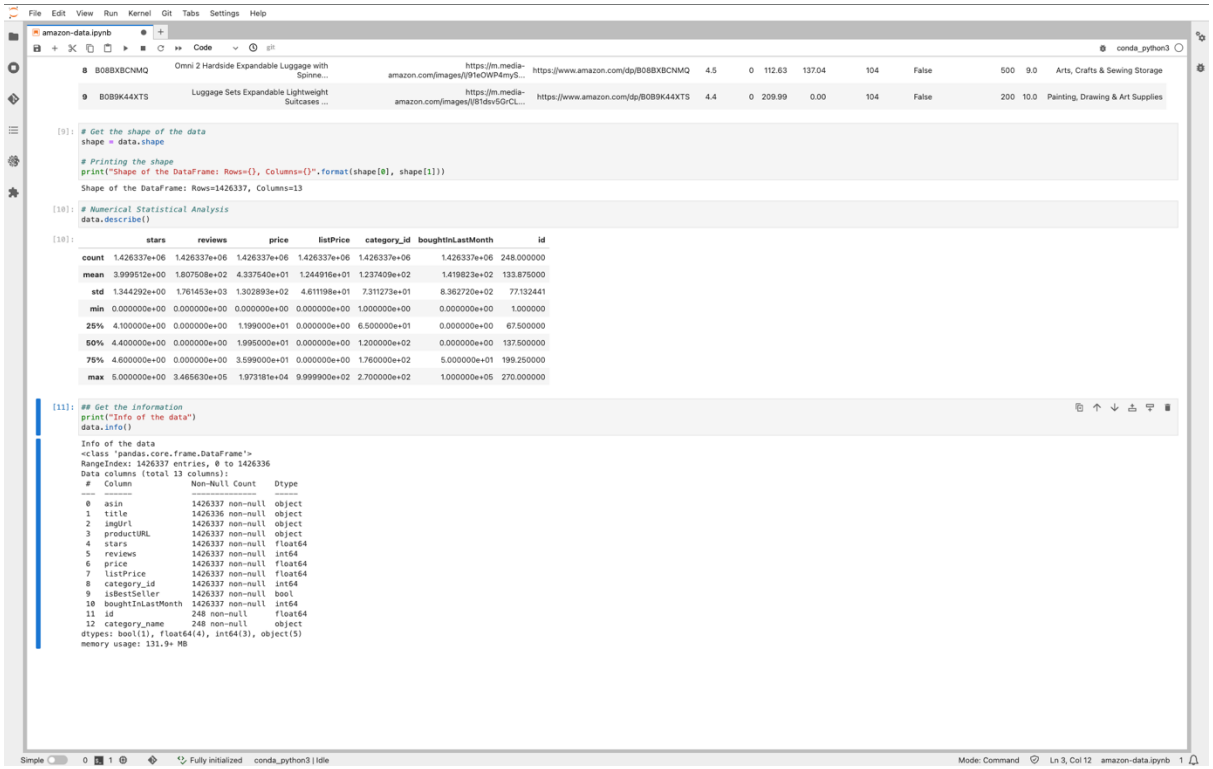

Services
Search
[Option+S]
N. Virginia
Amazon SageMaker
Success! Your notebook instance is being created.
Open the notebook instance when status is InService and open a template notebook to get started.
View details
Getting started
Studio
Studio Lab
Canvas
RStudio
TensorBoard
Profiler
Admin configurations
Domains
Role manager
Images
Lifecycle configurations
SageMaker dashboard
Search
JumpStart
Foundation models
Computer vision models
Natural language processing models
Governance
HyperPod Clusters
Ground Truth
Notebook
Notebook instances
Git repositories
Processing
Training
Inference
Edge Manager
Augmented AI
AWS Marketplace
Amazon SageMaker > Notebook instances
Notebook instances Info
Actions
Create notebook instance
Search notebook instances
Name
Instance
Creation time
Status
Actions
amazon-data-train
ml.m4.xlarge
12/3/2023, 11:57:04 AM
InService
Open Jupyter | Open JupyterLab

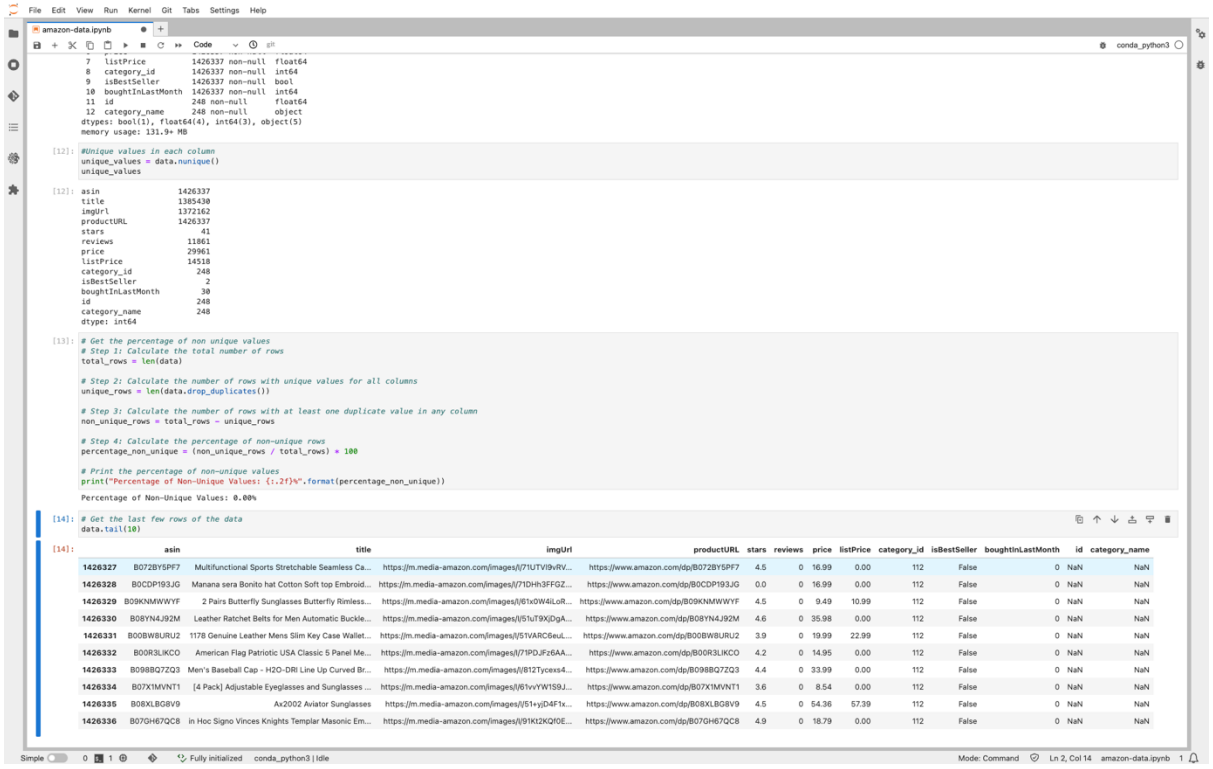

File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
amazon-data.ipynb
seconds ago
amazon-data.ipynb
Code
conda_python3
Would you like to receive official Jupyter news?
Please read the privacy policy.
Open privacy policy
Yes
No

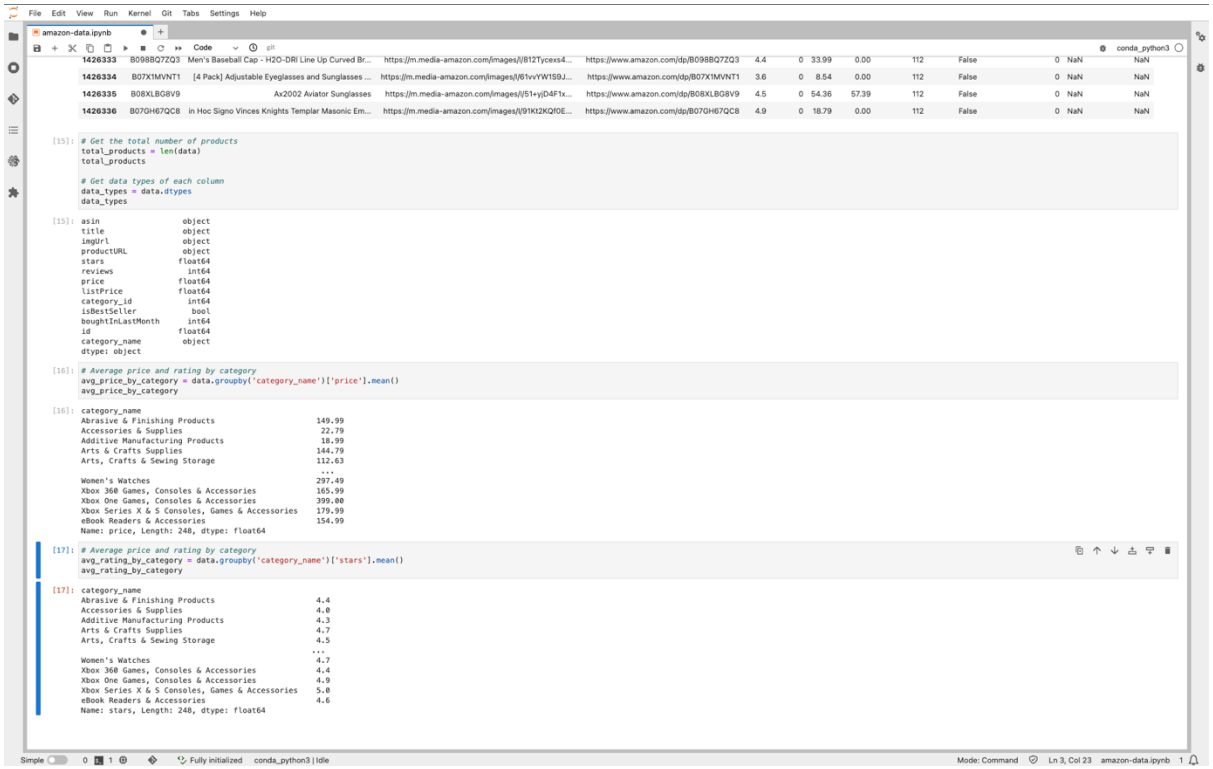


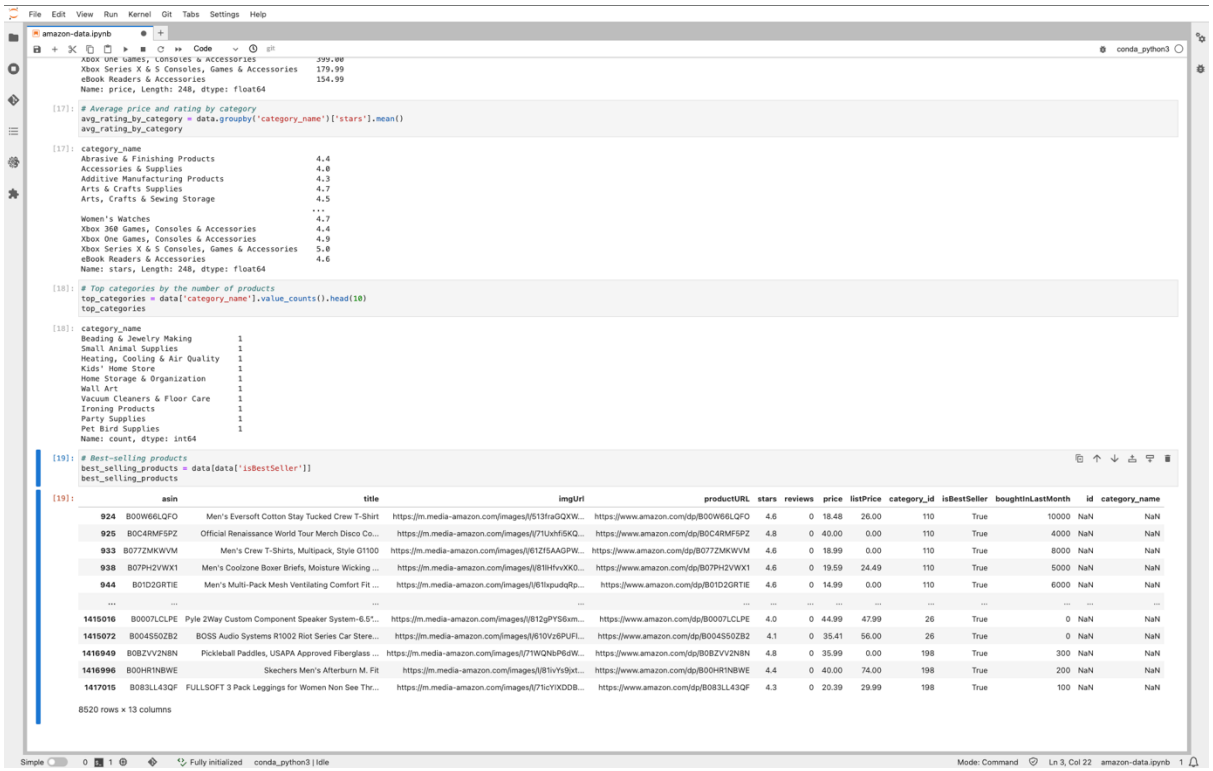


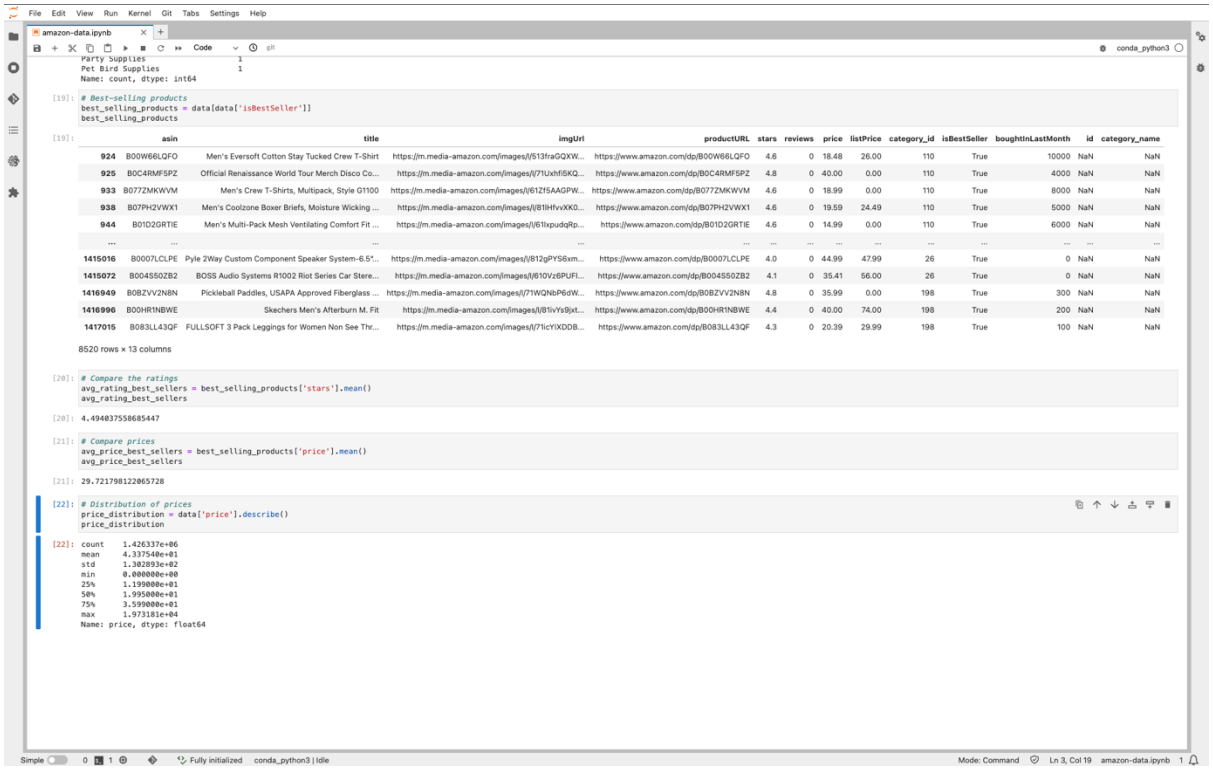
	asin	title	imgUrl	productURL	stars	reviews	price	listPrice	category_id	isBestSeller	boughtInLastMonth	id	category_name
0	B014TMV5YE	Sion Softside Expandable Roller Luggage, Black...	https://m.media-amazon.com/images/I/81SdLQKYiY...	https://www.amazon.com/dp/B014TMV5YE	4.5	0	139.99	0.00	104	False	2000	1.0	Beading & Jewelry Making
1	B07GDLCQXV	Luggage Sets Expandable PC+ABS Durable Suitcas...	https://m.media-amazon.com/images/I/81bQlm7vf6...	https://www.amazon.com/dp/B07GDLCQXV	4.5	0	169.99	209.99	104	False	1000	2.0	Fabric Decorating
2	B07XSCCZYG	Platinum Elite Softside Expandable Checked Lug...	https://m.media-amazon.com/images/I/71EA35zvJB...	https://www.amazon.com/dp/B07XSCCZYG	4.6	0	365.49	429.99	104	False	300	3.0	Knitting & Crochet Supplies
3	B08MVFKGJM	Freeform Hardside Expandable with Double Spinn...	https://m.media-amazon.com/images/I/91k6NYLQyI...	https://www.amazon.com/dp/B08MVFKGJM	4.6	0	291.59	354.37	104	False	400	4.0	Printmaking Supplies
4	B01DJLKZBA	Winfield 2 Hardside Expandable Luggage with Sp...	https://m.media-amazon.com/images/I/61NJoaZcPS...	https://www.amazon.com/dp/B01DJLKZBA	4.5	0	174.99	309.99	104	False	400	5.0	Scrapbooking & Stamping Supplies
5	B07XSCD2R4	Maxlite 5 Softside Expandable Luggage with 4 S...	https://m.media-amazon.com/images/I/61LnBNsSBS...	https://www.amazon.com/dp/B07XSCD2R4	4.5	0	144.49	0.00	104	False	500	6.0	Sewing Products
6	B07MXF4G8K	Hard Shell Carry on Luggage Airline Approved, ...	https://m.media-amazon.com/images/I/71CghLYrnA...	https://www.amazon.com/dp/B07MXF4G8K	4.5	0	169.99	0.00	104	False	400	7.0	Craft & Hobby Fabric
7	B07H515VCZ	Maxporter II 30" Hardside Spinner Trunk Luggag...	https://m.media-amazon.com/images/I/81f3h+YHOX...	https://www.amazon.com/dp/B07H515VCZ	4.5	0	299.99	0.00	104	False	100	8.0	Needlework Supplies
8	B08BXBCNMQ	Omni 2 Hardside Expandable Luggage with Spinne...	https://m.media-amazon.com/images/I/91eOWP4myS...	https://www.amazon.com/dp/B08BXBCNMQ	4.5	0	112.63	137.04	104	False	500	9.0	Arts, Crafts & Sewing Storage
9	B0B9K44XTS	Luggage Sets Expandable Lightweight Suitcases ...	https://m.media-amazon.com/images/I/81dsv5GrCL...	https://www.amazon.com/dp/B0B9K44XTS	4.4	0	209.99	0.00	104	False	200	10.0	Painting, Drawing & Art Supplies

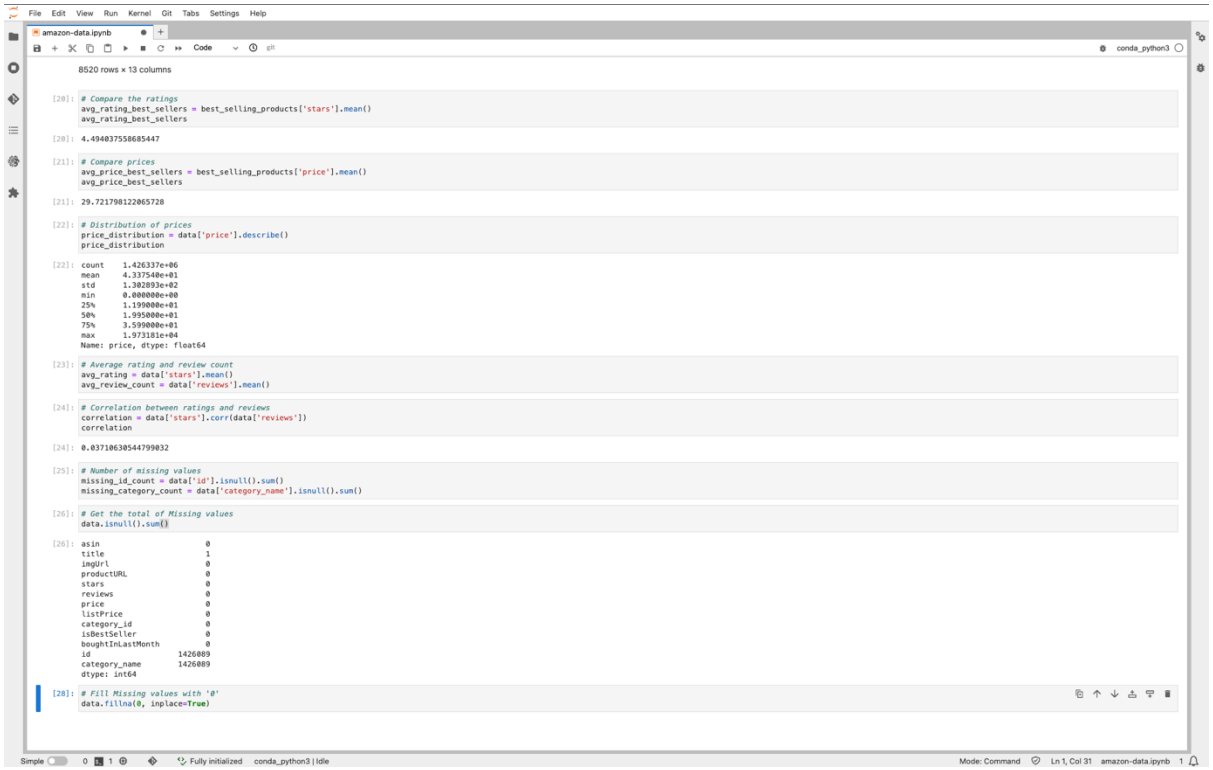

File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
8 B08BXBCNMQ Omni 2 Hardside Expandable Luggage with Spinne... https://m.media-amazon.com/images/I/91eOWP4myS... https://www.amazon.com/dp/B08BXBCNMQ 4.5 0 112.63 137.04 104 False 500 9.0 Arts, Crafts & Sewing Storage
9 B0B9K44XTS Luggage Sets Expandable Lightweight Suitcases ... https://m.media-amazon.com/images/I/81dsv5GrCL... https://www.amazon.com/dp/B0B9K44XTS 4.4 0 209.99 0.00 104 False 200 10.0 Painting, Drawing & Art Supplies
[9]: # Get the shape of the data
shape = data.shape
# Printing the shape
print("Shape of the DataFrame: Rows={}, Columns={}".format(shape[0], shape[1]))
Shape of the DataFrame: Rows=1426337, Columns=13
[10]: # Numerical Statistical Analysis
data.describe()
[10]:
stars reviews price listPrice category_id boughtInLastMonth id
count 1.426337e+06 1.426337e+06 1.426337e+06 1.426337e+06 1.426337e+06 1.426337e+06 248.000000
mean 3.999512e+00 1.807508e+02 4.337540e+01 1.244916e+01 1.237409e+02 1.419823e+02 133.875000
std 1.344292e+00 1.761453e+03 1.302893e+02 4.611198e+01 7.311273e+01 8.362720e+02 77.132441
min 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 1.000000e+00 0.000000e+00 1.000000
25% 4.100000e+00 0.000000e+00 1.199000e+01 0.000000e+00 6.500000e+01 0.000000e+00 67.500000
50% 4.400000e+00 0.000000e+00 1.995000e+01 0.000000e+00 1.200000e+02 0.000000e+00 137.500000
75% 4.600000e+00 0.000000e+00 3.599000e+01 0.000000e+00 1.760000e+02 5.000000e+01 199.250000
max 5.000000e+00 3.465630e+05 1.973181e+04 9.999900e+02 2.700000e+02 1.000000e+05 270.000000
[11]: ## Get the information
print("Info of the data")
data.info()
Info of the data
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1426337 entries, 0 to 1426336
Data columns (total 13 columns):
Column Non-Null Count Dtype
0 asin 1426337 non-null object
1 title 1426336 non-null object
2 imgUrl 1426337 non-null object
3 productURL 1426337 non-null object
4 stars 1426337 non-null float64
5 reviews 1426337 non-null int64
6 price 1426337 non-null float64
7 listPrice 1426337 non-null float64
8 category_id 1426337 non-null int64
9 isBestSeller 1426337 non-null bool
10 boughtInLastMonth 1426337 non-null int64
11 id 248 non-null float64
12 category_name 248 non-null object
dtypes: bool(1), float64(4), int64(3), object(5)
memory usage: 131.9+ MB
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 3, Col 12 amazon-data.ipynb 1


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
7 listPrice 1426337 non-null float64
8 category_id 1426337 non-null int64
9 isBestSeller 1426337 non-null bool
10 boughtInLastMonth 1426337 non-null int64
11 id 248 non-null float64
12 category_name 248 non-null object
dtypes: bool(1), float64(4), int64(3), object(5)
memory usage: 131.9+ MB
[12]: #Unique values in each column
unique_values = data.nunique()
unique_values
[12]:
asin 1426337
title 1385430
imgUrl 1372162
productURL 1426337
stars 41
reviews 11861
price 29961
listPrice 14518
category_id 248
isBestSeller 2
boughtInLastMonth 30
id 248
category_name 248
dtype: int64
[13]: # Get the percentage of non unique values
# Step 1: Calculate the total number of rows
total_rows = len(data)
# Step 2: Calculate the number of rows with unique values for all columns
unique_rows = len(data.drop_duplicates())
# Step 3: Calculate the number of rows with at least one duplicate value in any column
non_unique_rows = total_rows - unique_rows
# Step 4: Calculate the percentage of non-unique rows
percentage_non_unique = (non_unique_rows / total_rows) * 100
# Print the percentage of non-unique values
print("Percentage of Non-Unique Values: {:.2f}%".format(percentage_non_unique))
Percentage of Non-Unique Values: 0.00%
[14]: # Get the last few rows of the data
data.tail(10)
[14]:
asin title imgUrl productURL stars reviews price listPrice category_id isBestSeller boughtInLastMonth id category_name
1426327 B072BY5PF7 Multifunctional Sports Stretchable Seamless Ca... https://m.media-amazon.com/images/I/71UTVI9vRV... https://www.amazon.com/dp/B072BY5PF7 4.5 0 16.99 0.00 112 False 0 NaN NaN
1426328 B0CDP193JG Manana sera Bonito hat Cotton Soft top Embroid... https://m.media-amazon.com/images/I/71DHh3FFGZ... https://www.amazon.com/dp/B0CDP193JG 0.0 0 16.99 0.00 112 False 0 NaN NaN
1426329 B09KNMWWYF 2 Pairs Butterfly Sunglasses Butterfly Rimless... https://m.media-amazon.com/images/I/61x0W4iLoR... https://www.amazon.com/dp/B09KNMWWYF 4.5 0 9.49 10.99 112 False 0 NaN NaN
1426330 B08YN4J92M Leather Ratchet Belts for Men Automatic Buckle... https://m.media-amazon.com/images/I/51uT9XjDgA... https://www.amazon.com/dp/B08YN4J92M 4.6 0 35.98 0.00 112 False 0 NaN NaN
1426331 B00BW8URU2 1178 Genuine Leather Mens Slim Key Case Wallet... https://m.media-amazon.com/images/I/51VARC6euL... https://www.amazon.com/dp/B00BW8URU2 3.9 0 19.99 22.99 112 False 0 NaN NaN
1426332 B00R3LIKCO American Flag Patriotic USA Classic 5 Panel Me... https://m.media-amazon.com/images/I/71PDJFz6AA... https://www.amazon.com/dp/B00R3LIKCO 4.2 0 14.95 0.00 112 False 0 NaN NaN
1426333 B098BQ7ZQ3 Men's Baseball Cap - H2O-DRI Line Up Curved Br... https://m.media-amazon.com/images/I/812Tycexs4... https://www.amazon.com/dp/B098BQ7ZQ3 4.4 0 33.99 0.00 112 False 0 NaN NaN
1426334 B07X1MVNT1 [4 Pack] Adjustable Eyeglasses and Sunglasses ... https://m.media-amazon.com/images/I/61vvYW1S9J... https://www.amazon.com/dp/B07X1MVNT1 3.6 0 8.54 0.00 112 False 0 NaN NaN
1426335 B08XLBG8V9 Ax2002 Aviator Sunglasses https://m.media-amazon.com/images/I/51+yjD4F1x... https://www.amazon.com/dp/B08XLBG8V9 4.5 0 54.36 57.39 112 False 0 NaN NaN
1426336 B07GH67QC8 in Hoc Signo Vinces Knights Templar Masonic Em... https://m.media-amazon.com/images/I/91KI2KQf0E... https://www.amazon.com/dp/B07GH67QC8 4.9 0 18.79 0.00 112 False 0 NaN NaN
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 2, Col 14 amazon-data.ipynb 1


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
1426333 B098BQ7ZQ3 Men's Baseball Cap - H2O-DRI Line Up Curved Br... https://m.media-amazon.com/images/I/812Tycexs4... https://www.amazon.com/dp/B098BQ7ZQ3 4.4 0 33.99 0.00 112 False 0 NaN NaN
1426334 B07X1MVNT1 [4 Pack] Adjustable Eyeglasses and Sunglasses ... https://m.media-amazon.com/images/I/61vvYW1S9J... https://www.amazon.com/dp/B07X1MVNT1 3.6 0 8.54 0.00 112 False 0 NaN NaN
1426335 B08XLBG8V9 Ax2002 Aviator Sunglasses https://m.media-amazon.com/images/I/51+yjD4F1x... https://www.amazon.com/dp/B08XLBG8V9 4.5 0 54.36 57.39 112 False 0 NaN NaN
1426336 B07GH67QC8 in Hoc Signo Vinces Knights Templar Masonic Em... https://m.media-amazon.com/images/I/91K1t2KQf0E... https://www.amazon.com/dp/B07GH67QC8 4.9 0 18.79 0.00 112 False 0 NaN NaN
[15]:
# Get the total number of products
total_products = len(data)
total_products
# Get data types of each column
data_types = data.dtypes
data_types
[15]:
asin object
title object
imgUrl object
productURL object
stars float64
reviews int64
price float64
listPrice float64
category_id int64
isBestSeller bool
boughtInLastMonth int64
id float64
category_name object
dtype: object
[16]:
# Average price and rating by category
avg_price_by_category = data.groupby('category_name')['price'].mean()
avg_price_by_category
[16]:
category_name
Abrasive & Finishing Products 149.99
Accessories & Supplies 22.79
Additive Manufacturing Products 18.99
Arts & Crafts Supplies 144.79
Arts, Crafts & Sewing Storage 112.63
...
Women's Watches 297.49
Xbox 360 Games, Consoles & Accessories 165.99
Xbox One Games, Consoles & Accessories 399.00
Xbox Series X & S Consoles, Games & Accessories 179.99
eBook Readers & Accessories 154.99
Name: price, Length: 248, dtype: float64
[17]:
# Average price and rating by category
avg_rating_by_category = data.groupby('category_name')['stars'].mean()
avg_rating_by_category
[17]:
category_name
Abrasive & Finishing Products 4.4
Accessories & Supplies 4.0
Additive Manufacturing Products 4.3
Arts & Crafts Supplies 4.7
Arts, Crafts & Sewing Storage 4.5
...
Women's Watches 4.7
Xbox 360 Games, Consoles & Accessories 4.4
Xbox One Games, Consoles & Accessories 4.9
Xbox Series X & S Consoles, Games & Accessories 5.0
eBook Readers & Accessories 4.6
Name: stars, Length: 248, dtype: float64
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 3, Col 23 amazon-data.ipynb 1


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
Xbox One Games, Consoles & Accessories 399.00
Xbox Series X & S Consoles, Games & Accessories 179.99
eBook Readers & Accessories 154.99
Name: price, Length: 248, dtype: float64
[17]:
# Average price and rating by category
avg_rating_by_category = data.groupby('category_name')['stars'].mean()
avg_rating_by_category
[17]:
category_name
Abrasive & Finishing Products 4.4
Accessories & Supplies 4.0
Additive Manufacturing Products 4.3
Arts & Crafts Supplies 4.7
Arts, Crafts & Sewing Storage 4.5
...
Women's Watches 4.7
Xbox 360 Games, Consoles & Accessories 4.4
Xbox One Games, Consoles & Accessories 4.9
Xbox Series X & S Consoles, Games & Accessories 5.0
eBook Readers & Accessories 4.6
Name: stars, Length: 248, dtype: float64
[18]:
# Top categories by the number of products
top_categories = data['category_name'].value_counts().head(10)
top_categories
[18]:
category_name
Beading & Jewelry Making 1
Small Animal Supplies 1
Heating, Cooling & Air Quality 1
Kids' Home Store 1
Home Storage & Organization 1
Wall Art 1
Vacuum Cleaners & Floor Care 1
Ironing Products 1
Party Supplies 1
Pet Bird Supplies 1
Name: count, dtype: int64
[19]:
# Best-selling products
best_selling_products = data[data['isBestSeller']]
best_selling_products
[19]:
asin title imgUrl productURL stars reviews price listPrice category_id isBestSeller boughtInLastMonth id category_name
924 B00W66LQFO Men's Eversoft Cotton Stay Tucked Crew T-Shirt https://m.media-amazon.com/images/I/513fraGQXW... https://www.amazon.com/dp/B00W66LQFO 4.6 0 18.48 26.00 110 True 10000 NaN NaN
925 B0C4RMF5PZ Official Renaissance World Tour Merch Disco Co... https://m.media-amazon.com/images/I/71Uxhfi5KQ... https://www.amazon.com/dp/B0C4RMF5PZ 4.8 0 40.00 0.00 110 True 4000 NaN NaN
933 B077ZMKWVM Men's Crew T-Shirts, Multipack, Style G1100 https://m.media-amazon.com/images/I/61Zf5AAGPW... https://www.amazon.com/dp/B077ZMKWVM 4.6 0 18.99 0.00 110 True 8000 NaN NaN
938 B07PH2VWX1 Men's Coolzone Boxer Briefs, Moisture Wicking ... https://m.media-amazon.com/images/I/81lHfvvXK0... https://www.amazon.com/dp/B07PH2VWX1 4.6 0 19.59 24.49 110 True 5000 NaN NaN
944 B01D2GRTIE Men's Multi-Pack Mesh Ventilating Comfort Fit ... https://m.media-amazon.com/images/I/61IxpudqRp... https://www.amazon.com/dp/B01D2GRTIE 4.6 0 14.99 0.00 110 True 6000 NaN NaN
... ...
1415016 B0007LCLPE Pyle 2Way Custom Component Speaker System-6.5"... https://m.media-amazon.com/images/I/812gPYS6xm... https://www.amazon.com/dp/B0007LCLPE 4.0 0 44.99 47.99 26 True 0 NaN NaN
1415072 B004S50ZB2 BOSS Audio Systems R1002 Riot Series Car Stere... https://m.media-amazon.com/images/I/610Vz6PUFI... https://www.amazon.com/dp/B004S50ZB2 4.1 0 35.41 56.00 26 True 0 NaN NaN
1416949 B0BZVV2N8N Pickleball Paddles, USAPA Approved Fiberglass ... https://m.media-amazon.com/images/I/71WQNbP6dW... https://www.amazon.com/dp/B0BZVV2N8N 4.8 0 35.99 0.00 198 True 300 NaN NaN
1416996 B00HR1NBWE Skechers Men's Afterburn M. Fit https://m.media-amazon.com/images/I/81ivYs9jxt... https://www.amazon.com/dp/B00HR1NBWE 4.4 0 40.00 74.00 198 True 200 NaN NaN
1417015 B083LL43QF FULLSOFT 3 Pack Leggings for Women Non See Thr... https://m.media-amazon.com/images/I/71icYiXDDB... https://www.amazon.com/dp/B083LL43QF 4.3 0 20.39 29.99 198 True 100 NaN NaN
8520 rows × 13 columns
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 3, Col 22 amazon-data.ipynb 1


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
Party Supplies 1
Pet Bird Supplies 1
Name: count, dtype: int64
[19]: # Best-selling products
best_selling_products = data[data['isBestSeller']]
best_selling_products
[19]:
asin title imgUrl productURL stars reviews price listPrice category_id isBestSeller boughtInLastMonth id category_name
924 B00W66LQFO Men's Eversoft Cotton Stay Tucked Crew T-Shirt https://m.media-amazon.com/images/I/513fraQQXW... https://www.amazon.com/dp/B00W66LQFO 4.6 0 18.48 26.00 110 True 10000 NaN NaN
925 B0C4RMF5PZ Official Renaissance World Tour Merch Disco Co... https://m.media-amazon.com/images/I/71Uxhfi5KQ... https://www.amazon.com/dp/B0C4RMF5PZ 4.8 0 40.00 0.00 110 True 4000 NaN NaN
933 B077ZMKWVM Men's Crew T-Shirts, Multipack, Style G1100 https://m.media-amazon.com/images/I/61Zf5AAGPW... https://www.amazon.com/dp/B077ZMKWVM 4.6 0 18.99 0.00 110 True 8000 NaN NaN
938 B07PH2VWX1 Men's Coolzone Boxer Briefs, Moisture Wicking ... https://m.media-amazon.com/images/I/81lHfvvXK0... https://www.amazon.com/dp/B07PH2VWX1 4.6 0 19.59 24.49 110 True 5000 NaN NaN
944 B01D2GRTIE Men's Multi-Pack Mesh Ventilating Comfort Fit ... https://m.media-amazon.com/images/I/61lxpudqRp... https://www.amazon.com/dp/B01D2GRTIE 4.6 0 14.99 0.00 110 True 6000 NaN NaN
... ...
1415016 B0007CLCPE Pyle 2Way Custom Component Speaker System-6.5"... https://m.media-amazon.com/images/I/812gPYS6xm... https://www.amazon.com/dp/B0007CLCPE 4.0 0 44.99 47.99 26 True 0 NaN NaN
1415072 B004S50ZB2 BOSS Audio Systems R1002 Riot Series Car Stere... https://m.media-amazon.com/images/I/610Vz6PUFI... https://www.amazon.com/dp/B004S50ZB2 4.1 0 35.41 56.00 26 True 0 NaN NaN
1416949 B0BZVV2N8N Pickleball Paddles, USAPA Approved Fiberglass ... https://m.media-amazon.com/images/I/71WQNbP6dW... https://www.amazon.com/dp/B0BZVV2N8N 4.8 0 35.99 0.00 198 True 300 NaN NaN
1416996 B00HR1NBWE Skechers Men's Afterburn M. Fit https://m.media-amazon.com/images/I/81ivYs9jxt... https://www.amazon.com/dp/B00HR1NBWE 4.4 0 40.00 74.00 198 True 200 NaN NaN
1417015 B083LL43QF FULLSOFT 3 Pack Leggings for Women Non See Thr... https://m.media-amazon.com/images/I/71icYiXDDB... https://www.amazon.com/dp/B083LL43QF 4.3 0 20.39 29.99 198 True 100 NaN NaN
8520 rows × 13 columns
[20]: # Compare the ratings
avg_rating_best_sellers = best_selling_products['stars'].mean()
avg_rating_best_sellers
[20]: 4.494037558685447
[21]: # Compare prices
avg_price_best_sellers = best_selling_products['price'].mean()
avg_price_best_sellers
[21]: 29.721798122065728
[22]: # Distribution of prices
price_distribution = data['price'].describe()
price_distribution
[22]: count 1.426337e+06
mean 4.337540e+01
std 1.302893e+02
min 0.000000e+00
25% 1.199000e+01
50% 1.995000e+01
75% 3.599000e+01
max 1.973181e+04
Name: price, dtype: float64
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 3, Col 19 amazon-data.ipynb 1


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
8520 rows × 13 columns
[20]: # Compare the ratings
avg_rating_best_sellers = best_selling_products['stars'].mean()
avg_rating_best_sellers
[20]: 4.494037558685447
[21]: # Compare prices
avg_price_best_sellers = best_selling_products['price'].mean()
avg_price_best_sellers
[21]: 29.721798122065728
[22]: # Distribution of prices
price_distribution = data['price'].describe()
price_distribution
[22]: count 1.426337e+06
mean 4.337540e+01
std 1.302893e+02
min 0.000000e+00
25% 1.199000e+01
50% 1.995000e+01
75% 3.599000e+01
max 1.973181e+04
Name: price, dtype: float64
[23]: # Average rating and review count
avg_rating = data['stars'].mean()
avg_review_count = data['reviews'].mean()
[24]: # Correlation between ratings and reviews
correlation = data['stars'].corr(data['reviews'])
correlation
[24]: 0.03710630544799032
[25]: # Number of missing values
missing_id_count = data['id'].isnull().sum()
missing_category_count = data['category_name'].isnull().sum()
[26]: # Get the total of Missing values
data.isnull().sum()
[26]: asin 0
title 1
imgUrl 0
productURL 0
stars 0
reviews 0
price 0
listPrice 0
category_id 0
isBestSeller 0
boughtInLastMonth 0
id 1426089
category_name 1426089
dtype: int64
[28]: # Fill Missing values with '0'
data.fillna(0, inplace=True)
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 1, Col 31 amazon-data.ipynb 1

File Edit View Run Kernel Git Tabs Settings Help

amazon-data.ipynb

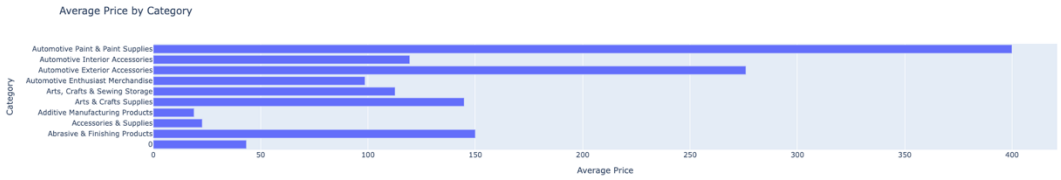
Code

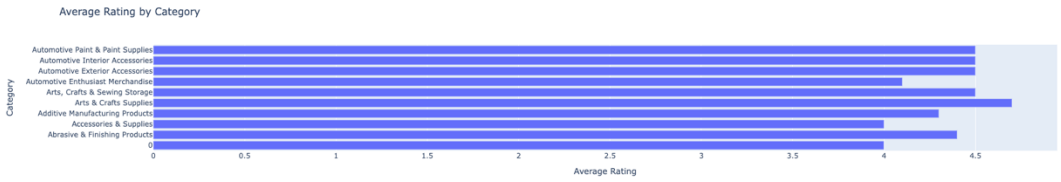
conda_python3

```
data.isnull().sum().sum()
```

[29]:
```
0
```

[31]:
```
import plotly.express as px

# . Calculate the average price and rating for products in each category
avg_price_by_category = data.groupby('category_name')['price'].mean().reset_index()
avg_price_by_category.columns = ['Category', 'Average Price']

avg_rating_by_category = data.groupby('category_name')['stars'].mean().reset_index()
avg_rating_by_category.columns = ['Category', 'Average Rating']


# . Create an interactive bar chart for average prices
fig2 = px.bar(avg_price_by_category.head(10), x='Average Price', y='Category', orientation='h',
              title='Average Price by Category',
              labels={'Average Price': 'Average Price'})

# . Create an interactive bar chart for average ratings
fig3 = px.bar(avg_rating_by_category.head(10), x='Average Rating', y='Category', orientation='h',
              title='Average Rating by Category',
              labels={'Average Rating': 'Average Rating'})

# 6. Display the interactive charts
fig2.show()
fig3.show()
```





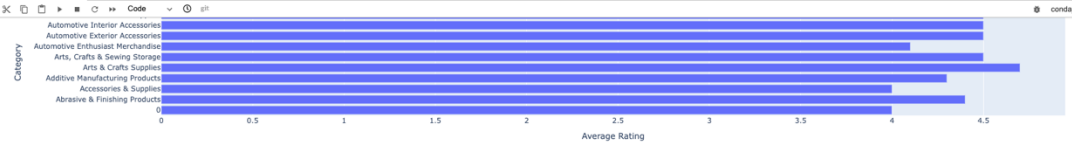
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 2, Col 1 amazon-data.ipynb 1

File Edit View Run Kernel Git Tabs Settings Help

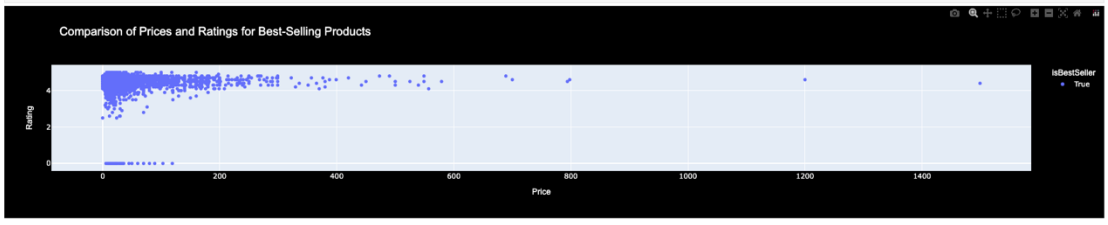
amazon-data.ipynb

Code

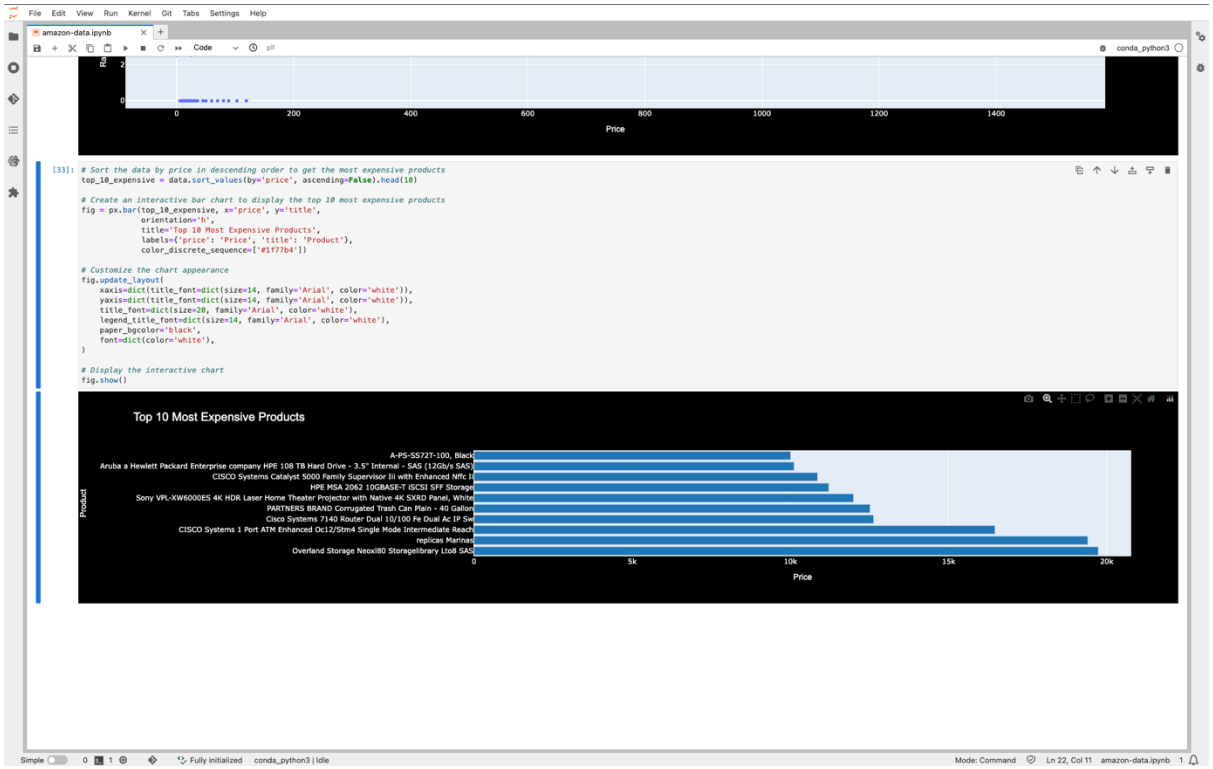
conda_python3

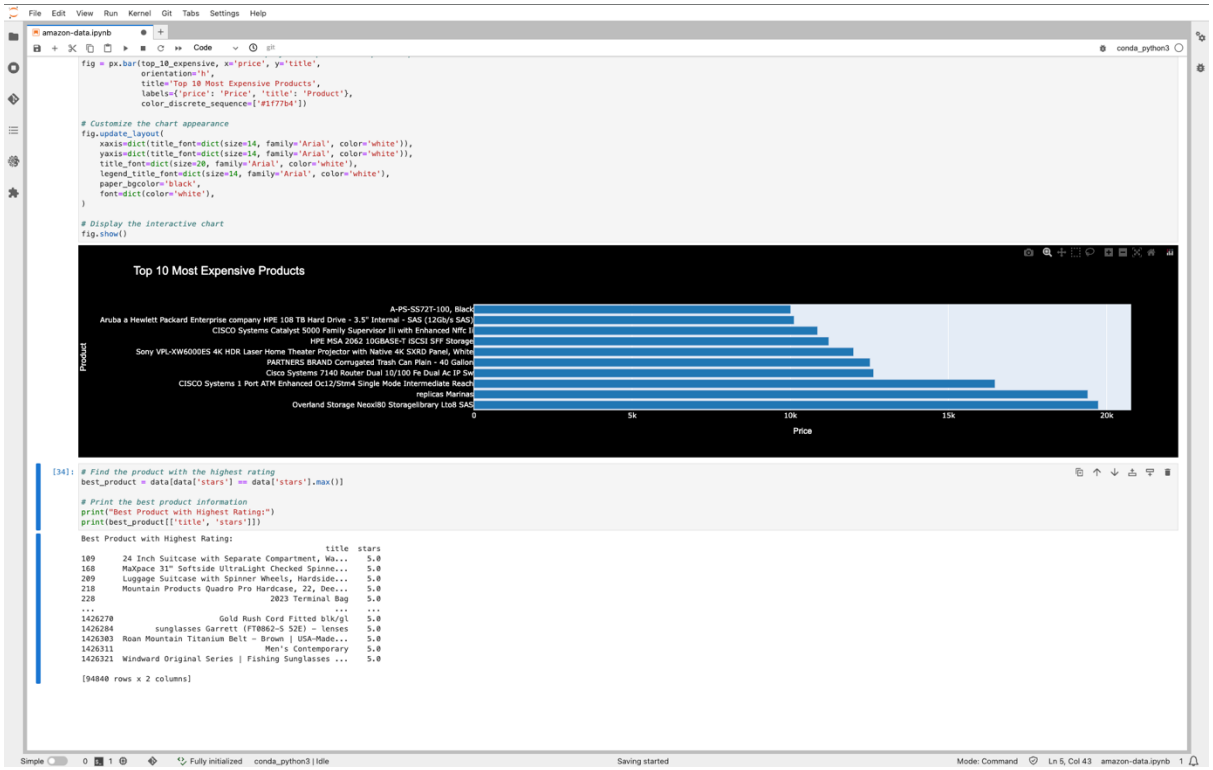


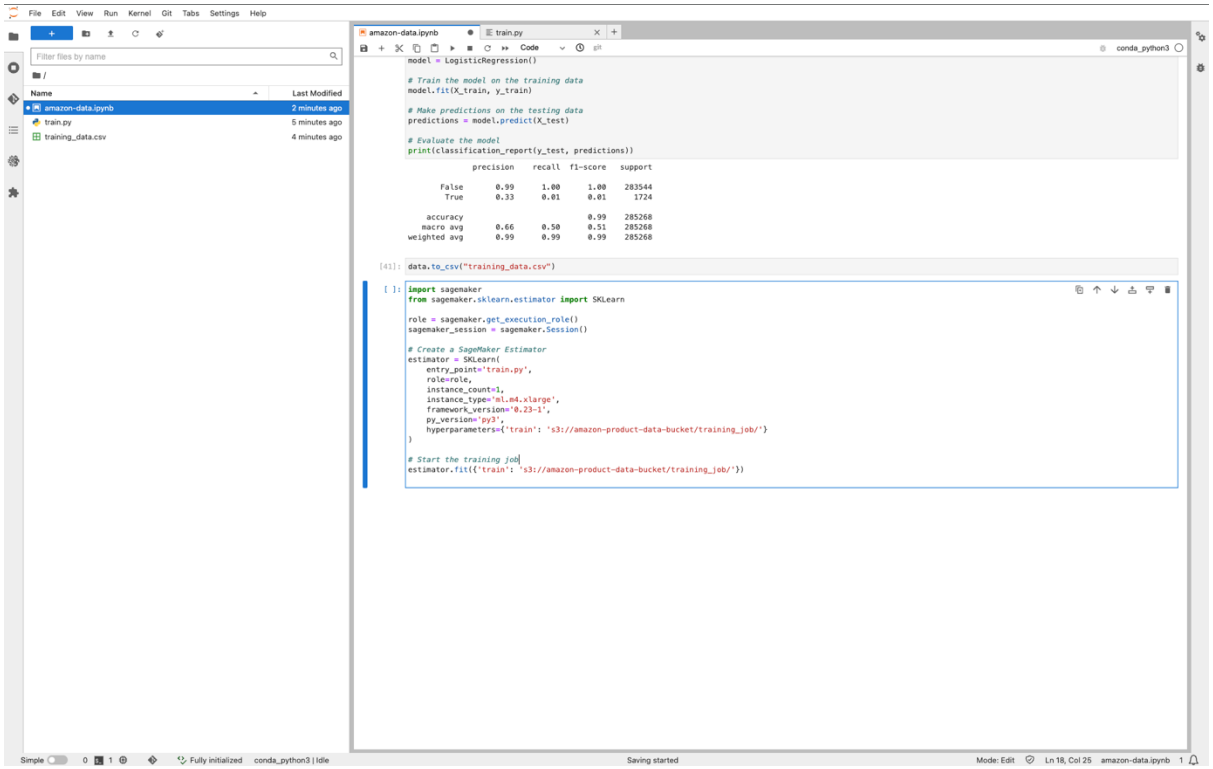
[32]:
```
# Filter the DataFrame to get best-selling products
best_selling_products = data[data['isBestSeller']]

# Create a scatter plot to compare prices and ratings
fig = px.scatter(best_selling_products, x='price', y='stars', color='isBestSeller',
                 title='Comparison of Prices and Ratings for Best-Selling Products',
                 labels={'price': 'Price', 'stars': 'Rating'},
                 hover_data=['title'])

# Customize the chart
fig.update_layout(
    xaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
    yaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
    title_font=dict(size=20, family='Arial', color='white'),
    legend_title_font=dict(size=14, family='Arial', color='white'),
    paper_bgcolor='black',
    font=dict(color='white'),
)

# Display the interactive chart
fig.show()
```

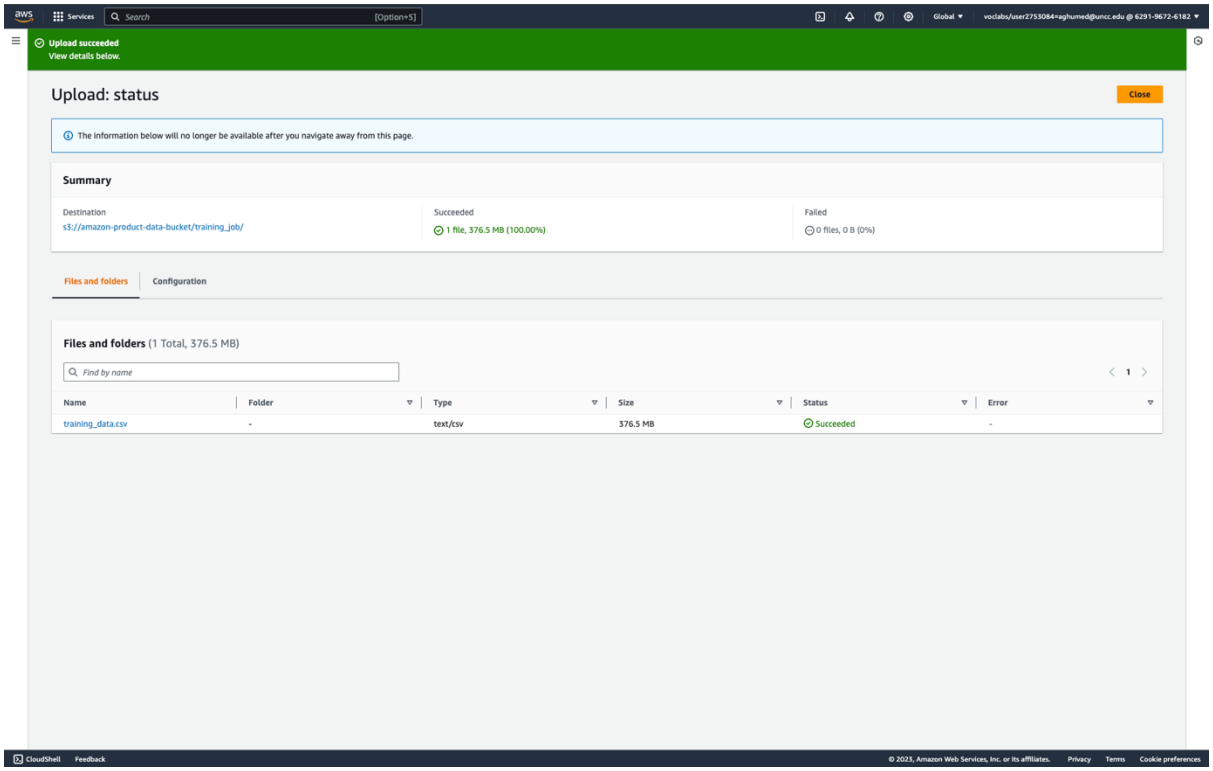


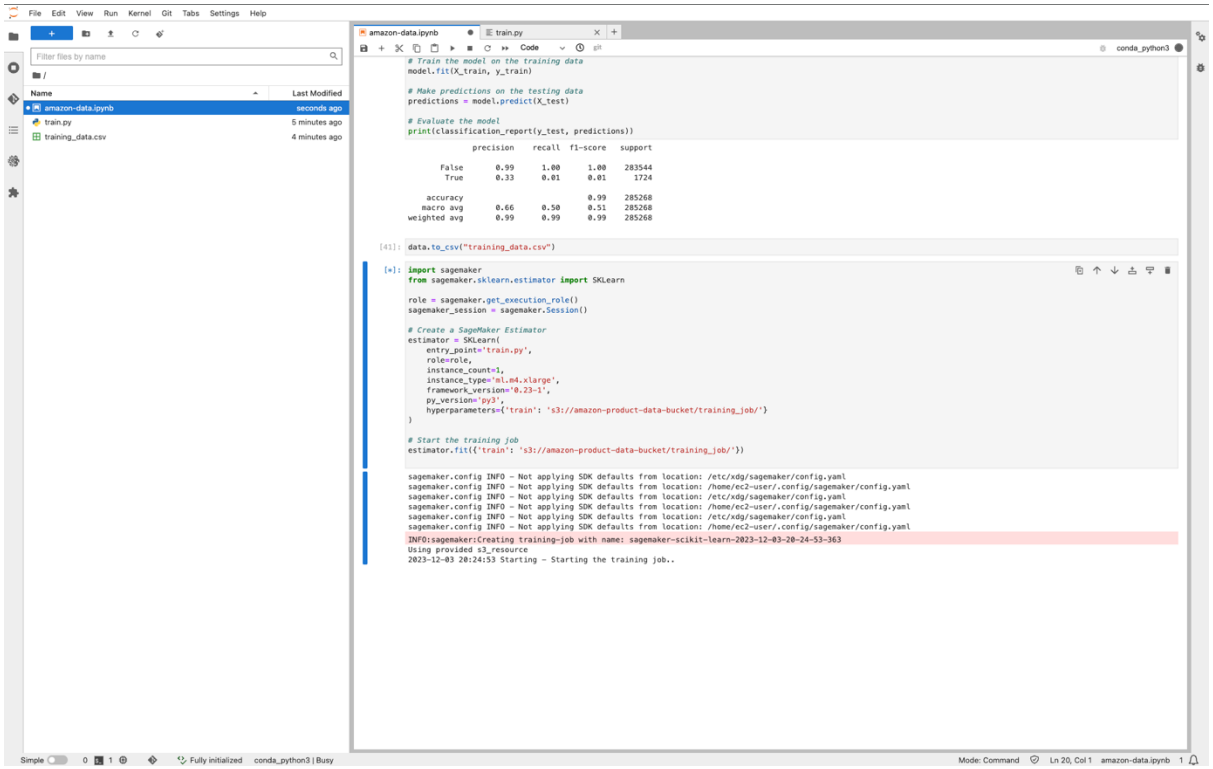
Simple 0 1 Fully initialized conda_python3 | Idle Mode: Command Ln 21, Col 11 amazon-data.ipynb 1

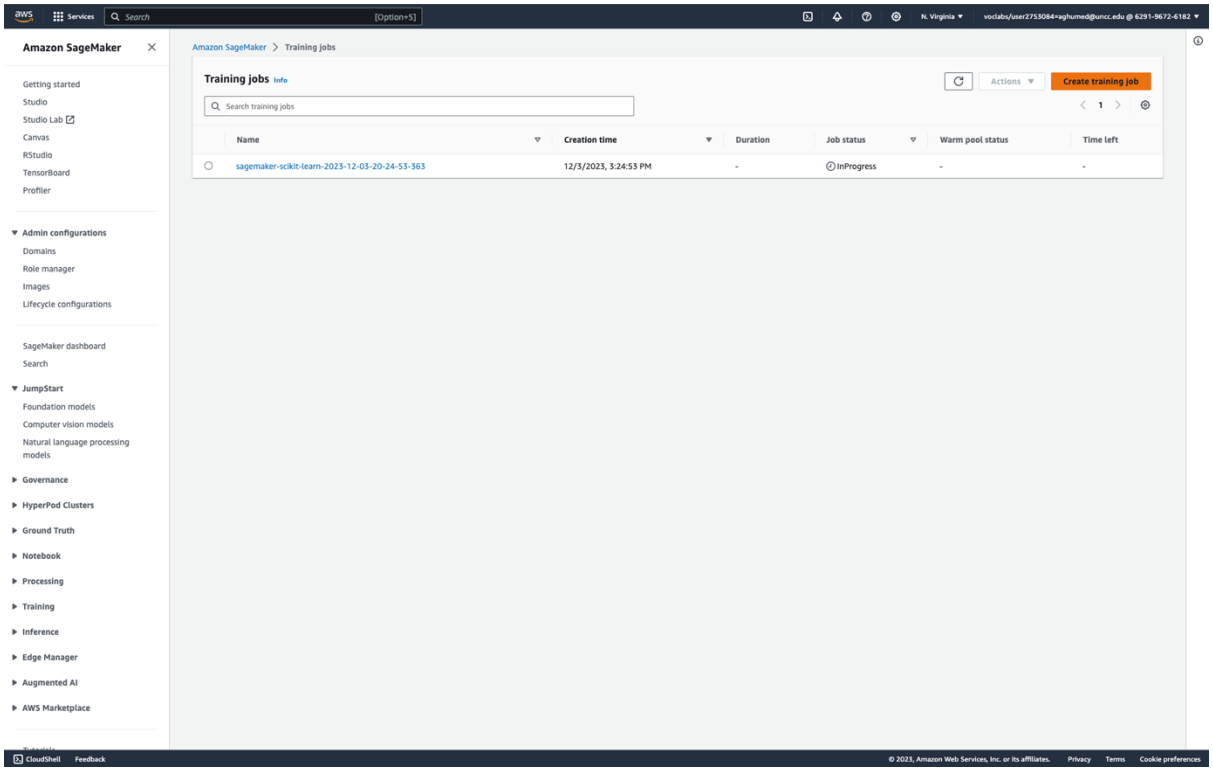

File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
Rating
2
0
0
200
400
600
800
1000
1200
1400
Price
[33]:
# Sort the data by price in descending order to get the most expensive products
top_10_expensive = data.sort_values(by='price', ascending=False).head(10)
# Create an interactive bar chart to display the top 10 most expensive products
fig = px.bar(top_10_expensive, x='price', y='title',
orientation='h',
title='Top 10 Most Expensive Products',
labels={'price': 'Price', 'title': 'Product'},
color_discrete_sequence=['#1f77b4'])
# Customize the chart appearance
fig.update_layout(
xaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
yaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
title_font=dict(size=20, family='Arial', color='white'),
legend_title_font=dict(size=14, family='Arial', color='white'),
paper_bgcolor='black',
font=dict(color='white'),
)
# Display the interactive chart
fig.show()
Top 10 Most Expensive Products
Product
A-PS-SS72T-100, Black
Aruba a Hewlett Packard Enterprise company HPE 108 TB Hard Drive - 3.5" Internal - SAS (12Gb/s SAS)
CISCO Systems Catalyst 5000 Family Supervisor Iii with Enhanced Nffc I
HPE MSA 2062 10GBASE-T iSCSI SFF Storage
Sony VPL-XW6000ES 4K HDR Laser Home Theater Projector with Native 4K SXRD Panel, White
PARTNERS BRAND Corrugated Trash Can Plain - 40 Gallon
Cisco Systems 7140 Router Dual 10/100 Fe Dual Ac IP Sw
CISCO Systems 1 Port ATM Enhanced Oc12/Stm4 Single Mode Intermediate Reach
replicas Marinas
Overland Storage Neoxl80 Storagelibrary Lto8 SAS
0
5k
10k
15k
20k
Price
Simple
Fully initialized
conda_python3 | Idle
Mode: Command
Ln 22, Col 11
amazon-data.ipynb

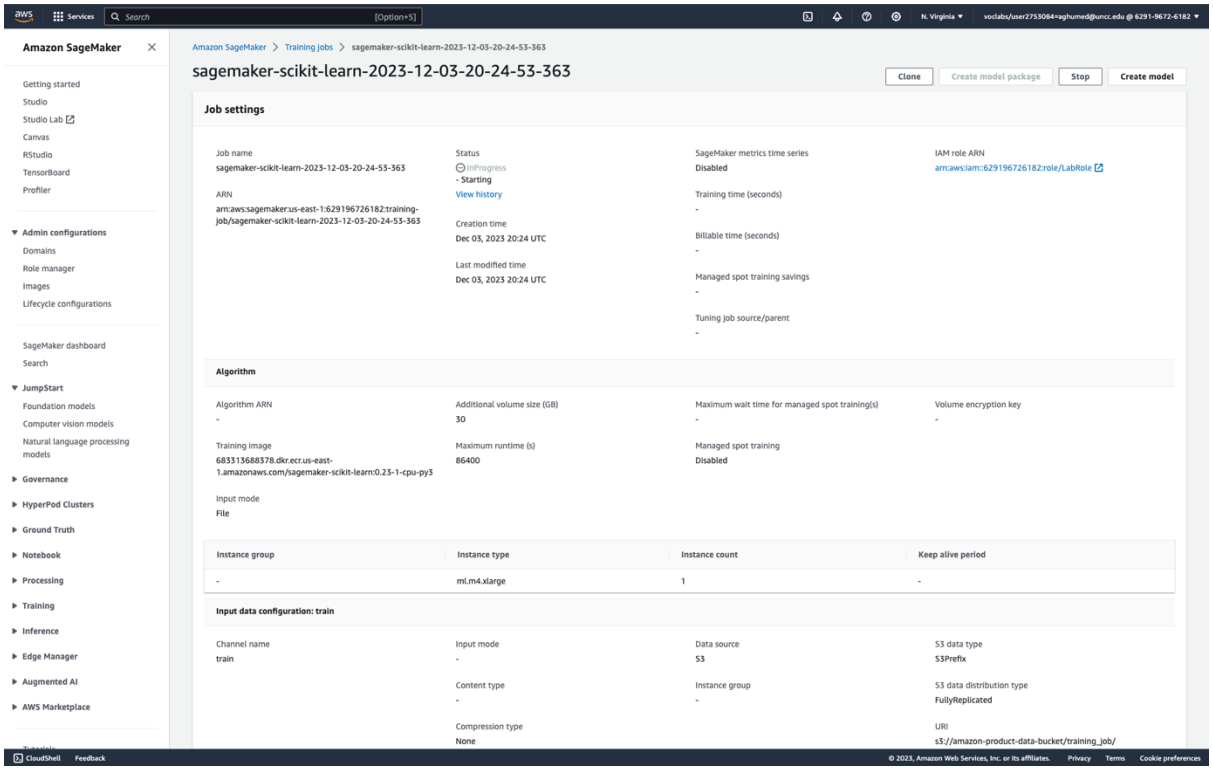

File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
fig = px.bar(top_10_expensive, x='price', y='title',
orientation='h',
title='Top 10 Most Expensive Products',
labels={'price': 'Price', 'title': 'Product'},
color_discrete_sequence=['#1f77b4'])
# Customize the chart appearance
fig.update_layout(
xaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
yaxis=dict(title_font=dict(size=14, family='Arial', color='white')),
title_font=dict(size=20, family='Arial', color='white'),
legend_title_font=dict(size=14, family='Arial', color='white'),
paper_bgcolor='black',
font=dict(color='white'),
)
# Display the interactive chart
fig.show()
Top 10 Most Expensive Products
Product
A-PS-SS72T-100, Black
Aruba a Hewlett Packard Enterprise company HPE 108 TB Hard Drive - 3.5" Internal - SAS (12Gb/s SAS)
CISCO Systems Catalyst 5000 Family Supervisor Iii with Enhanced Nffc I
HPE MSA 2062 10GBASE-T iSCSI SFF Storage
Sony VPL-XW6000ES 4K HDR Laser Home Theater Projector with Native 4K SXRD Panel, White
PARTNERS BRAND Corrugated Trash Can Plain - 40 Gallon
Cisco Systems 7140 Router Dual 10/100 Fe Dual Ac IP Sw
CISCO Systems 1 Port ATM Enhanced Oc12/Stm4 Single Mode Intermediate Reach
replicas Marinas
Overland Storage Neoxl80 Storagelibrary Lto8 SAS
0
5k
10k
15k
20k
Price
[34]:
# Find the product with the highest rating
best_product = data[data['stars'] == data['stars'].max()]
# Print the best product information
print("Best Product with Highest Rating:")
print(best_product[['title', 'stars']])
Best Product with Highest Rating:
title stars
109 24 Inch Suitcase with Separate Compartment, Wa... 5.0
168 MaXpace 31" Softside UltraLight Checked Spinne... 5.0
209 Luggage Suitcase with Spinner Wheels, Hardside... 5.0
218 Mountain Products Quadro Pro Hardcase, 22, Dee... 5.0
228 2023 Terminal Bag 5.0
... ...
1426270 Gold Rush Cord Fitted blk/gl 5.0
1426284 sunglasses Garrett (FT0862-S 52E) - lenses 5.0
1426303 Roan Mountain Titanium Belt - Brown | USA-Made... 5.0
1426311 Men's Contemporary 5.0
1426321 Windward Original Series | Fishing Sunglasses ... 5.0
[94840 rows x 2 columns]
Simple
Fully initialized
conda_python3 | Idle
Saving started
Mode: Command
Ln 5, Col 43
amazon-data.ipynb

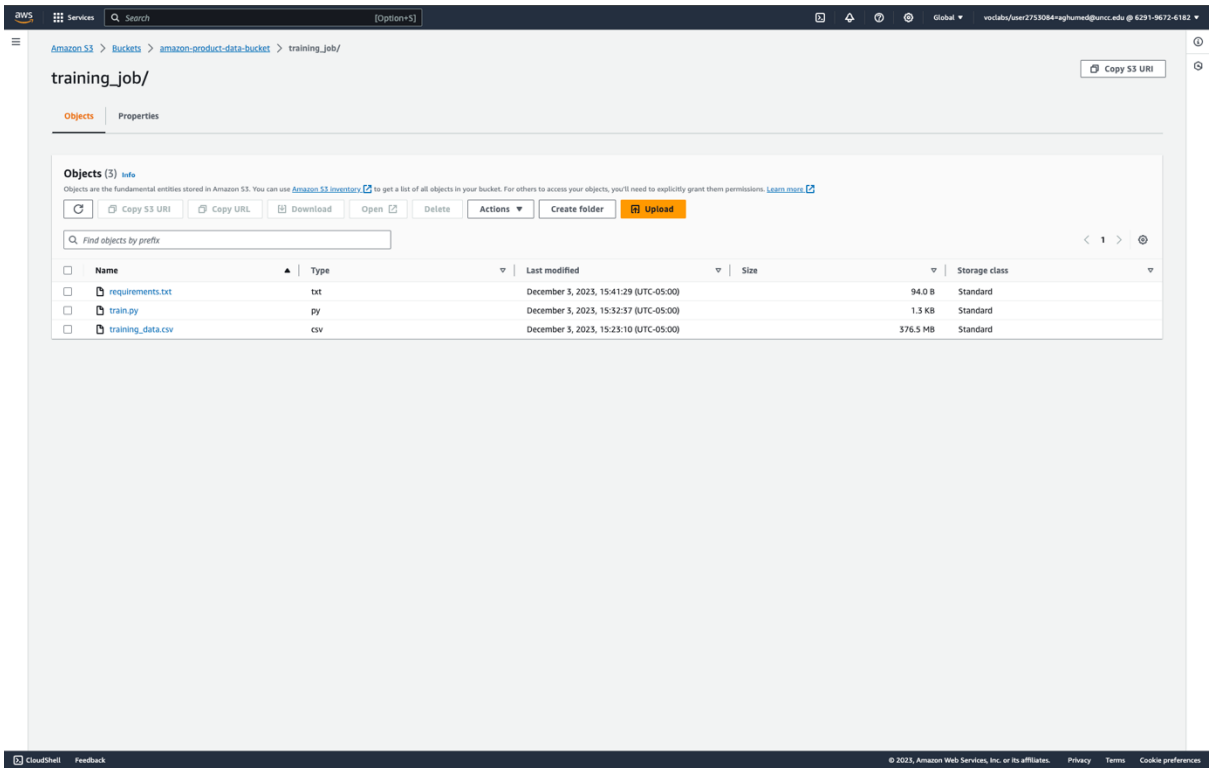

File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
amazon-data.ipynb
2 minutes ago
train.py
5 minutes ago
training_data.csv
4 minutes ago
amazon-data.ipynb
train.py
Code
conda_python3
model = LogisticRegression()
# Train the model on the training data
model.fit(X_train, y_train)
# Make predictions on the testing data
predictions = model.predict(X_test)
# Evaluate the model
print(classification_report(y_test, predictions))
precision recall f1-score support
False 0.99 1.00 1.00 283544
True 0.33 0.01 0.01 1724
accuracy 0.99 285268
macro avg 0.66 0.50 0.51 285268
weighted avg 0.99 0.99 0.99 285268
[41]: data.to_csv("training_data.csv")
[ ]: import sagemaker
from sagemaker.sklearn.estimator import SKLearn
role = sagemaker.get_execution_role()
sagemaker_session = sagemaker.Session()
# Create a SageMaker Estimator
estimator = SKLearn(
entry_point='train.py',
role=role,
instance_count=1,
instance_type='ml.m4.xlarge',
framework_version='0.23-1',
py_version='py3',
hyperparameters={'train': 's3://amazon-product-data-bucket/training_job/'}
)
# Start the training job
estimator.fit({'train': 's3://amazon-product-data-bucket/training_job/'})
Simple
Fully initialized
conda_python3 | Idle
Saving started
Mode: Edit
Ln 18, Col 25
amazon-data.ipynb

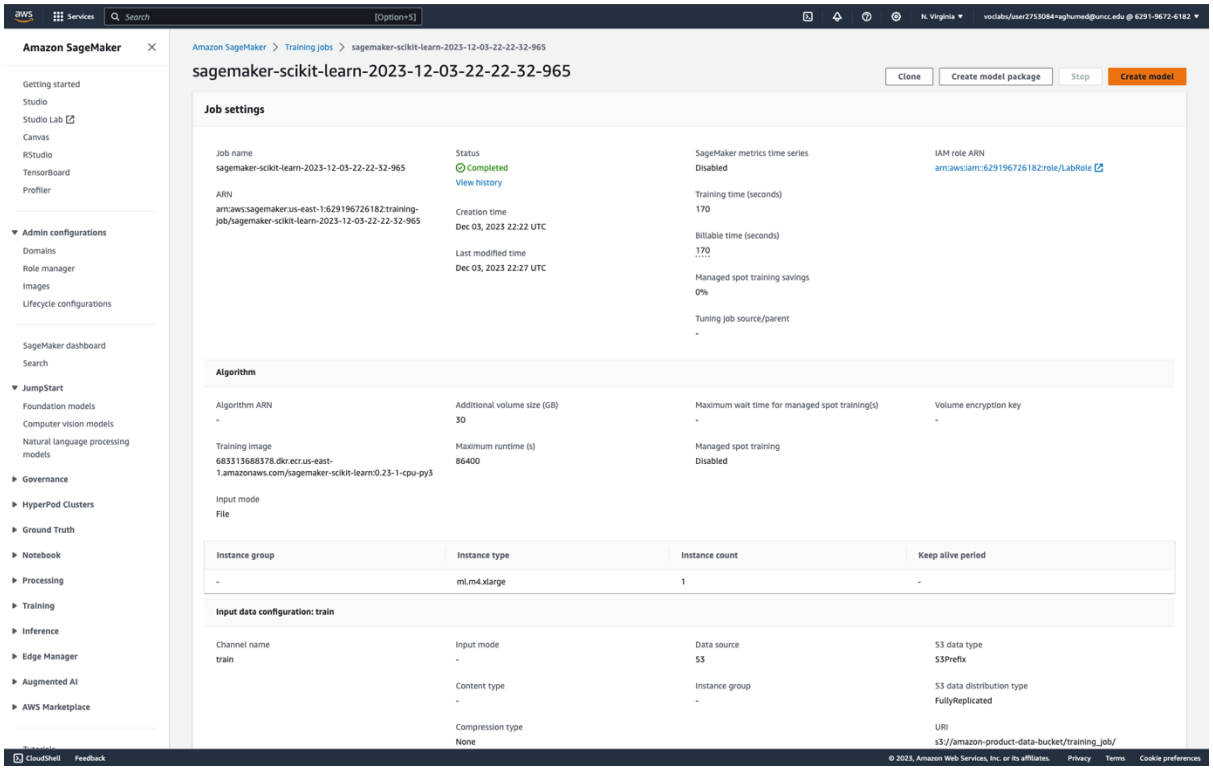

aws
Services
Search
[Option+S]
Global
voclabs/user2753084=aghumed@uncc.edu @ 6291-9672-6182
Upload succeeded
View details below.
Upload: status
Close
The information below will no longer be available after you navigate away from this page.
Summary
Destination
s3://amazon-product-data-bucket/training_job/
Succeeded
1 file, 376.5 MB (100.00%)
Failed
0 files, 0 B (0%)
Files and folders
Configuration
Files and folders (1 Total, 376.5 MB)
Find by name
Name
Folder
Type
Size
Status
Error
training_data.csv
-
text/csv
376.5 MB
Succeeded
-
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

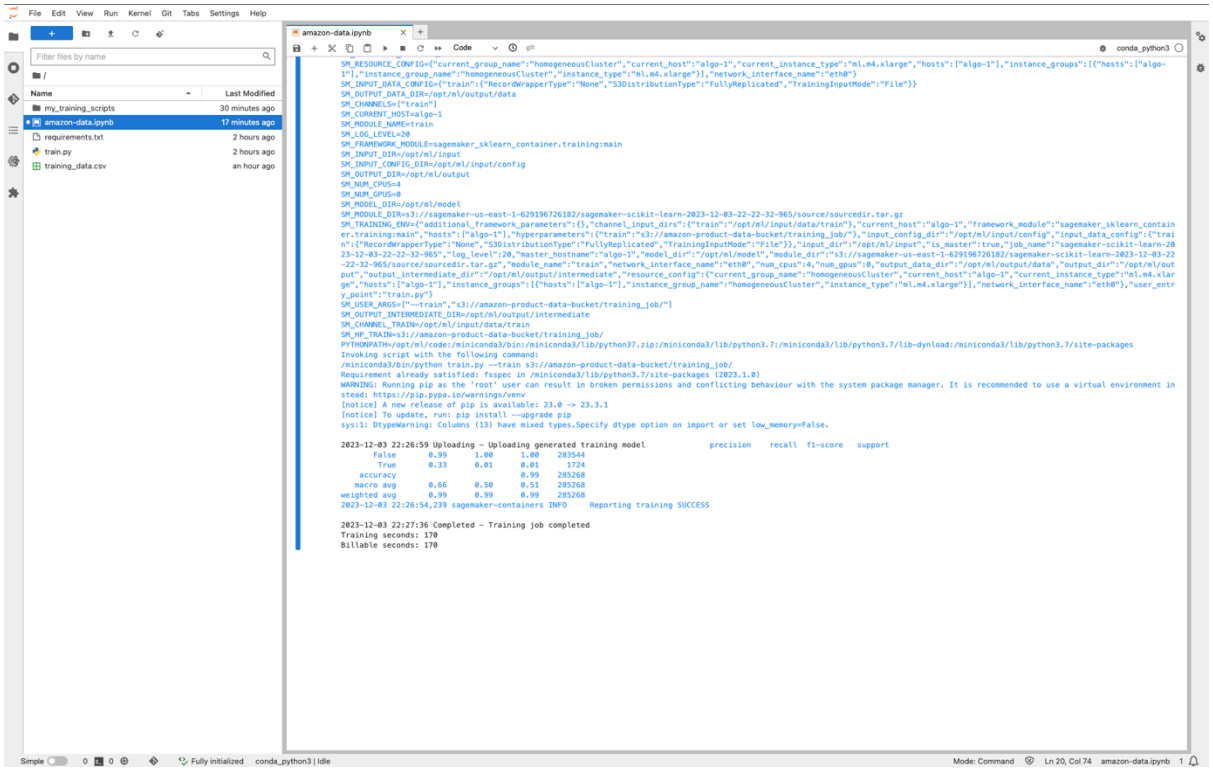

File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
amazon-data.ipynb
seconds ago
train.py
5 minutes ago
training_data.csv
4 minutes ago
amazon-data.ipynb
train.py
Code
conda_python3
# Train the model on the training data
model.fit(X_train, y_train)
# Make predictions on the testing data
predictions = model.predict(X_test)
# Evaluate the model
print(classification_report(y_test, predictions))
precision recall f1-score support
False 0.99 1.00 1.00 283544
True 0.33 0.01 0.01 1724
accuracy 0.99 285268
macro avg 0.66 0.50 0.51 285268
weighted avg 0.99 0.99 0.99 285268
[41]:
data.to_csv("training_data.csv")
[*]:
import sagemaker
from sagemaker.sklearn.estimator import SKLearn
role = sagemaker.get_execution_role()
sagemaker_session = sagemaker.Session()
# Create a SageMaker Estimator
estimator = SKLearn(
entry_point='train.py',
role=role,
instance_count=1,
instance_type='ml.m4.xlarge',
framework_version='0.23-1',
py_version='py3',
hyperparameters={'train': 's3://amazon-product-data-bucket/training_job/'}
)
# Start the training job
estimator.fit({'train': 's3://amazon-product-data-bucket/training_job/'})
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
INFO:sagemaker:Creating training-job with name: sagemaker-scikit-learn-2023-12-03-20-24-53-363
Using provided s3_resource
2023-12-03 20:24:53 Starting - Starting the training job..
Simple
Fully initialized
conda_python3 | Busy
Mode: Command
Ln 20, Col 1
amazon-data.ipynb


Services
Search
[Option+S]
N. Virginia
Amazon SageMaker
Getting started
Studio
Studio Lab
Canvas
RStudio
TensorBoard
Profiler
Admin configurations
Domains
Role manager
Images
Lifecycle configurations
SageMaker dashboard
Search
JumpStart
Foundation models
Computer vision models
Natural language processing models
Governance
HyperPod Clusters
Ground Truth
Notebook
Processing
Training
Inference
Edge Manager
Augmented AI
AWS Marketplace
Amazon SageMaker > Training jobs
Training jobs Info
Search training jobs
Actions
Create training job
Name
Creation time
Duration
Job status
Warm pool status
Time left
sagemaker-scikit-learn-2023-12-03-20-24-53-363
12/3/2023, 3:24:53 PM
-
InProgress
-
-
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon SageMaker
Training jobs
sagemaker-scikit-learn-2023-12-03-20-24-53-363
Clone
Create model package
Stop
Create model
Job settings
Job name
sagemaker-scikit-learn-2023-12-03-20-24-53-363
Status
InProgress
- Starting
View history
SageMaker metrics time series
Disabled
IAM role ARN
arn:aws:iam::629196726182:role/LabRole
ARN
arn:aws:sagemaker:us-east-1:629196726182:training-job/sagemaker-scikit-learn-2023-12-03-20-24-53-363
Creation time
Dec 03, 2023 20:24 UTC
Training time (seconds)
-
Billable time (seconds)
-
Last modified time
Dec 03, 2023 20:24 UTC
Managed spot training savings
-
Tuning job source/parent
-
Algorithm
Algorithm ARN
-
Additional volume size (GB)
30
Maximum wait time for managed spot training(s)
-
Volume encryption key
-
Training image
683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-scikit-learn:0.23-1-cpu-py3
Maximum runtime (s)
86400
Managed spot training
Disabled
Input mode
File
Instance group
Instance type
Instance count
Keep alive period
-
ml.m4.xlarge
1
-
Input data configuration: train
Channel name
train
Input mode
-
Data source
S3
S3 data type
S3Prefix
Content type
-
Instance group
-
S3 data distribution type
FullyReplicated
Compression type
None
URI
s3://amazon-product-data-bucket/training_job/
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon S3
Buckets
amazon-product-data-bucket
training_job/
training_job/
Copy S3 URI
Objects
Properties
Objects (3)
Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
requirements.txt
txt
December 3, 2023, 15:41:29 (UTC-05:00)
94.0 B
Standard
train.py
py
December 3, 2023, 15:32:37 (UTC-05:00)
1.3 KB
Standard
training_data.csv
csv
December 3, 2023, 15:23:10 (UTC-05:00)
376.5 MB
Standard
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon SageMaker
Amazon SageMaker > Training jobs > sagemaker-scikit-learn-2023-12-03-22-22-32-965
sagemaker-scikit-learn-2023-12-03-22-22-32-965
Clone
Create model package
Stop
Create model
Job settings
Job name
sagemaker-scikit-learn-2023-12-03-22-22-32-965
Status
Completed
View history
SageMaker metrics time series
Disabled
IAM role ARN
arn:aws:iam::629196726182:role/LabRole
ARN
arn:aws:sagemaker:us-east-1:629196726182:training-job/sagemaker-scikit-learn-2023-12-03-22-22-32-965
Creation time
Dec 03, 2023 22:22 UTC
Last modified time
Dec 03, 2023 22:27 UTC
Training time (seconds)
170
Billable time (seconds)
170
Managed spot training savings
0%
Tuning job source/parent
-
Algorithm
Algorithm ARN
-
Additional volume size (GB)
30
Maximum wait time for managed spot training(s)
-
Volume encryption key
-
Training image
683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-scikit-learn:0.23-1-cpu-py3
Maximum runtime (s)
86400
Managed spot training
Disabled
Input mode
File
Instance group
Instance type
Instance count
Keep alive period
-
ml.m4.xlarge
1
-
Input data configuration: train
Channel name
train
Input mode
-
Data source
S3
S3 data type
S3Prefix
Content type
-
Instance group
-
S3 data distribution type
FullyReplicated
Compression type
None
URI
s3://amazon-product-data-bucket/training_job/
CloudShell
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


File Edit View Run Kernel Git Tabs Settings Help
amazon-data.ipynb
Code
conda_python3
my_training_scripts
amazon-data.ipynb
requirements.txt
train.py
training_data.csv
2023-12-03 22:26:59 Uploading - Uploading generated training model
precision recall f1-score support
False 0.99 1.00 1.00 283544
True 0.33 0.01 0.01 1724
accuracy 0.99 285268
macro avg 0.66 0.50 0.51 285268
weighted avg 0.99 0.99 0.99 285268
2023-12-03 22:26:54,239 sagemaker-containers INFO Reporting training SUCCESS
2023-12-03 22:27:36 Completed - Training job completed
Training seconds: 170
Billable seconds: 170