Análise de Componentes Principais: Conjunto de dados de viagem do viajante

Kemelly Santos Sabino Breno Henrique Rey Lorenzo

June 6, 2023

1 Introdução

A análise de dados de viagem é uma ferramenta essencial para compreender os padrões e preferências dos viajantes. Neste artigo, adotaremos a técnica de Análise de Componentes Principais (PCA) como uma abordagem eficaz para explorar e extrair informações relevantes desse conjunto de dados.

O PCA é uma técnica estatística que nos permite reduzir a dimensionalidade dos dados, transformando um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas, conhecidas como componentes principais. Ao aplicar o PCA aos dados de viagem, seremos capazes de identificar os principais fatores que contribuem para a variabilidade nos padrões de viagem, como características demográficas dos viajantes e suas preferências de destino e acomodação.

Essa redução da dimensionalidade dos dados simplifica a análise, facilitando a identificação de tendências e insights relevantes. Além disso, ao eliminar variáveis redundantes ou menos significativas, podemos otimizar o desempenho de modelos preditivos ou algoritmos de aprendizado de máquina que possam ser aplicados posteriormente.

Por meio da aplicação do PCA, esperamos fornecer uma visão clara e abrangente das características e comportamentos dos viajantes, permitindo uma compreensão mais profunda do setor de viagens e turismo. Esses insights serão fundamentais para o desenvolvimento de estratégias de negócios personalizadas e bem-sucedidas, visando atender às demandas e expectativas dos viajantes modernos.

2 Objetivo

O objetivo deste artigo é utilizar a técnica de Análise de Componentes Principais (PCA) para explorar e extrair insights relevantes a partir do conjunto de dados de viagem. Pretendemos identificar os principais fatores que influenciam os padrões de viagem, como preferências de destino, duração da viagem, dados demográficos dos viajantes e detalhes sobre acomodação e transporte. Além disso, buscamos compreender as relações entre essas variáveis e visualizar

graficamente os padrões identificados. Ao aplicar o PCA, pretendemos reduzir a dimensionalidade do conjunto de dados, simplificando a análise e facilitando a identificação de tendências significativas. Com isso, pretendemos fornecer uma visão abrangente das características e comportamentos dos viajantes, permitindo uma compreensão mais profunda do setor de viagens e turismo. Esses insights serão úteis para empresas relacionadas a viagens, permitindo o desenvolvimento de estratégias de marketing personalizadas e a criação de pacotes de viagens que atendam às necessidades e preferências dos diferentes segmentos de mercado.

3 Desenvolvimento

Para realizar a análise dos componentes principais, utilizamos a biblioteca pandas para carregar os dados do conjunto de dados de viagem a partir de um arquivo Excel. Em seguida, são selecionadas as colunas relevantes para a análise de PCA, que são 'Destination' (destino), 'Accommodation type' (tipo de acomodação), 'Accommodation cost' (custo da acomodação) e 'Transportation cost' (custo do transporte).

Após selecionar as colunas, as variáveis categóricas são codificadas usando a técnica de one-hot encoding, e os valores ausentes são tratados utilizando a estratégia de substituição pela média. Em seguida, os dados são normalizados utilizando o StandardScaler.

É criado um objeto PCA e ajustado aos dados normalizados. O código imprime o tamanho da matriz de autovetores e o melhor autovalor, que corresponde à primeira componente principal. Também é identificada a linha do conjunto de dados que possui o melhor autovalor e é impressa apenas as colunas 'Destination' e 'Accommodation type' dessa linha.

Em seguida, o código obtém as categorias das variáveis categóricas para colorir os pontos nos gráficos de dispersão. O primeiro gráfico é um gráfico de dispersão em 2D que mostra os dados projetados nas duas primeiras componentes principais, onde cada ponto é colorido de acordo com a categoria do destino. O segundo gráfico é um gráfico de dispersão em 3D que mostra os dados projetados nas três primeiras componentes principais, também com cores indicando as categorias dos destinos.

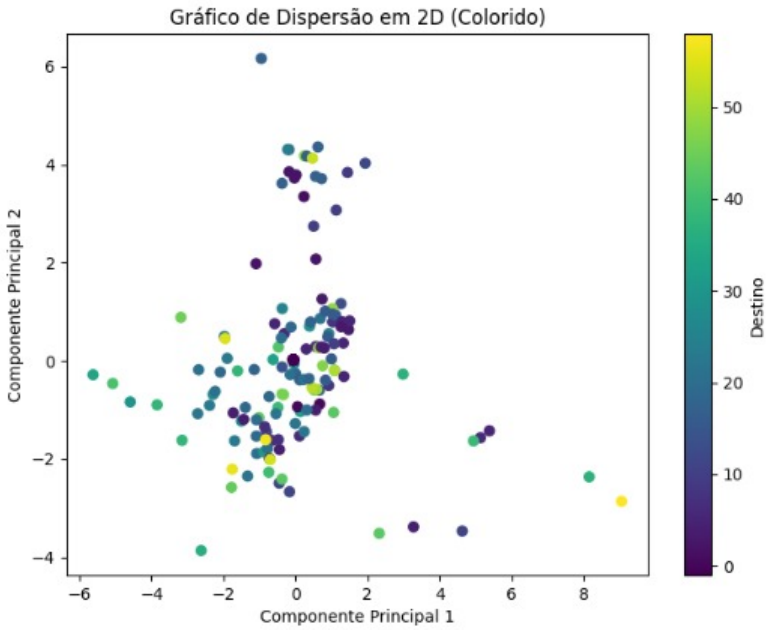
Essas visualizações permitem uma análise exploratória dos dados de viagem após a aplicação da técnica de PCA, fornecendo insights sobre os padrões e agrupamentos presentes nos dados.

```
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
from sklearn.decomposition import PCA
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer

# Carregar os dados do arquivo Excel
dados = pd.read_excel
(r"C:/Users/breno/Desktop/python2/Travel_details_dataset.xlsx.xltx")

# Selecionar as colunas relevantes para a análise de PCA
dados_selecionados = dados[['Destination', 'Accommodation type',
'Accommodation cost','Transportation cost']]

# Codificar as variáveis categóricas usando one-hot encoding
dados_codificados = pd.get_dummies(dados_selecionados)

# Tratar valores ausentes
imputer = SimpleImputer(strategy='mean')
dados_imputados = imputer.fit_transform(dados_codificados)

# Normalizar os dados
scaler = StandardScaler()
dados_normalizados = scaler.fit_transform(dados_imputados)

# Criar o objeto PCA e ajustá-lo aos dados normalizados
pca = PCA()
dados_transformados = pca.fit_transform(dados_normalizados)

# Imprimir o tamanho da matriz de autovetores
print("Tamanho da Matriz de Autovetores:", pca.components_.shape)

# Imprimir o melhor autovalor
melhor_autovalor = pca.explained_variance_[0]
print("Melhor Autovalor:", melhor_autovalor)

# Identificar a linha com o melhor autovalor
indice_melhor_autovalor = np.argmax(pca.components_[0])
linha_melhor_autovalor = dados.iloc[indice_melhor_autovalor]

# Imprimir a linha com o melhor autovalor (apenas as colunas
'Destination' e 'Accommodation type')
print("Linha com o Melhor Autovalor:")
print(linha_melhor_autovalor[['Destination', 'Accommodation type']])

# Obter a variável categórica para colorir os pontos
categorias = pd.factorize(dados_selecionados['Destination'])[0]

# Plotar o gráfico de dispersão em 2D com cores
```

```
fig = plt.figure(figsize=(8, 6))
ax = fig.add_subplot(111)
sc = ax.scatter(dados_transformados[:, 0], dados_transformados
[:, 1], c=categorias)
ax.set_xlabel('Componente Principal 1')
ax.set_ylabel('Componente Principal 2')
ax.set_title('Gráfico de Dispersão em 2D (Colorido)')
plt.colorbar(sc, label='Destino')
plt.show()

# Plotar o gráfico de dispersão em 3D com cores
fig = plt.figure(figsize=(8, 6))
ax = fig.add_subplot(111, projection='3d')
sc = ax.scatter(dados_transformados[:, 0], dados_transformados[:, 1],
dados_transformados[:, 2], c=categorias)
ax.set_xlabel('Componente Principal 1')
ax.set_ylabel('Componente Principal 2')
ax.set_zlabel('Componente Principal 3')
ax.set_title('Gráfico de Dispersão em 3D (Colorido)')
plt.colorbar(sc, label='Destino')
plt.show()
```

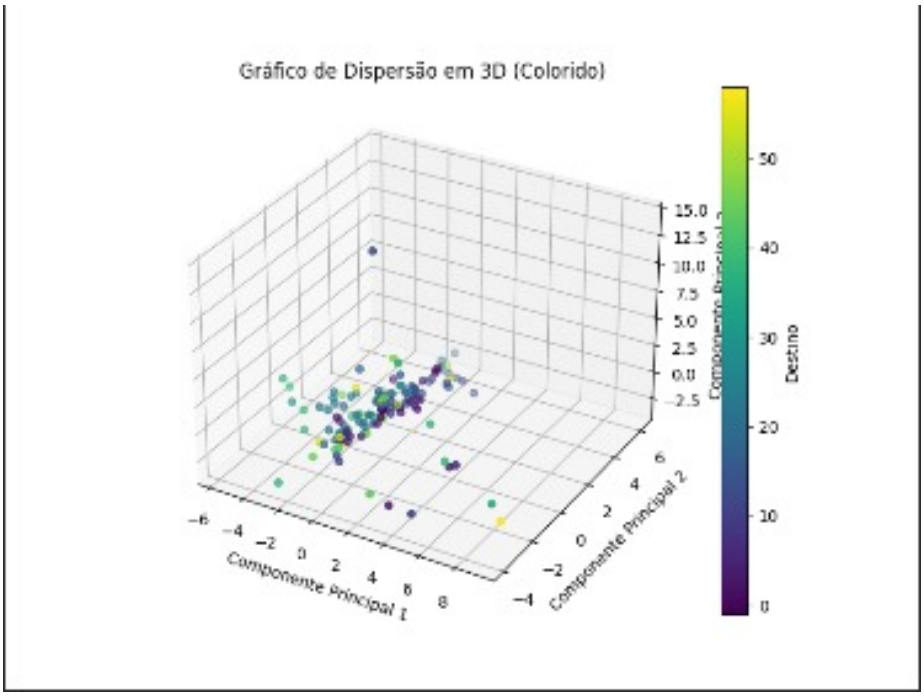
4 Análise



No gráfico de dispersão em 2D, podemos observar a distribuição dos destinos de viagem em relação às duas principais componentes principais obtidas pela técnica de Análise de Componentes Principais (PCA).

Cada ponto no gráfico representa uma viagem, e sua posição no plano bidimensional é determinada pelas projeções nas duas primeiras componentes principais. A cor dos pontos indica o destino da viagem. Ao observar o gráfico, podemos identificar padrões de agrupamento ou tendências na distribuição dos destinos.

A interpretação do gráfico de dispersão em 2D nos permite analisar as relações entre as variáveis selecionadas (destino, tipo de acomodação, custo de acomodação e custo de transporte) e entender como elas contribuem para a variação dos dados. Podemos observar se existem clusters específicos de destinos com base nas características das viagens, como tipo de acomodação e custos associados.



Com base nos dados fornecidos anteriormente, o gráfico de dispersão em 3D mostra a distribuição dos destinos de viagem em relação às três principais componentes principais obtidas pela técnica de Análise de Componentes Principais (PCA).

Cada ponto no gráfico representa uma viagem, e sua posição no espaço tridimensional é determinada pelas projeções nos três componentes principais. A cor dos pontos indica o destino da viagem. Ao observar o gráfico, podemos identificar agrupamentos de viagens semelhantes ou padrões distintos nas distribuições dos destinos.

A interpretação do gráfico de dispersão em 3D nos permite analisar as relações entre as variáveis selecionadas (destino, tipo de acomodação, custo de acomodação e custo de transporte) e entender como elas contribuem para a variação dos dados. Podemos observar se existem agrupamentos específicos de destinos com base nas características das viagens, como tipo de acomodação e custos associados.

Essa visualização tridimensional oferece uma perspectiva mais completa e detalhada dos padrões e associações nos dados de viagem, permitindo uma análise mais aprofundada das relações entre as variáveis e dos possíveis clusters ou tendências presentes.

5 Conclusão

Através da aplicação da técnica de Análise de Componentes Principais (PCA) nos dados de viagem, pudemos obter insights valiosos sobre os padrões, preferências e comportamentos dos viajantes.

Ao realizar a análise exploratória dos dados, identificamos as principais componentes que influenciam as escolhas e preferências dos viajantes, permitindo uma compreensão mais clara dos fatores que impulsionam o setor de turismo. Além disso, a utilização do PCA nos possibilitou visualizar graficamente as relações entre as variáveis, facilitando a interpretação dos resultados.

Através dos gráficos de dispersão em 2D e 3D, pudemos observar a distribuição dos destinos de viagem e identificar possíveis agrupamentos e tendências. A análise desses gráficos nos permitiu inferir relações entre as variáveis selecionadas, como tipo de acomodação, custo de acomodação e custo de transporte, e a escolha dos destinos. Essas informações são de grande importância para empresas relacionadas a viagens, como agências de turismo, que podem utilizar esses insights na criação de estratégias de marketing personalizadas e no desenvolvimento de pacotes de viagens adequados às necessidades e preferências dos viajantes.

Portanto, concluímos que a análise de dados de viagem, utilizando técnicas como o PCA, é uma abordagem eficaz para explorar e extrair informações relevantes sobre os padrões de viagem e preferências dos viajantes. Esses insights são essenciais para o setor de turismo, pois permitem o desenvolvimento de estratégias mais direcionadas, melhorando a oferta de serviços e produtos turísticos e atendendo às demandas dos viajantes modernos.