EDA (Exploratory Data Analysis) 탐색적 데이터 분석

웹 클립

EDA (Exploratory Data Analysis) 탐색적 데이터 분석

초코 오호힛 2018.11.12 15:36

1. EDA란?

1) 정의

수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정.한마디로 데이터를 분석하기 전에 그래프나 통계적인 방법으로 자료를 직관적으로 바라보는 과정이다.

2) 필요한 이유

첫째, 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 데이터에 대한 잠재적인 문제를 발견할 수 있다. 이를 통해, 본격적인 분석에 들어가기에 앞서 수집의사를 결정할 수 있다.

둘째, 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 미

쳐 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 세울 수 있다.

3) 과정

기본적인 출발점은 문제 정의 단계에서 세웠던 연구 질문과 가설을 바탕으로 분석 계획을 세우는 것이다. 분석 계획에는 어떤 속성 및 속성 간의 관계를 집중적으로 관찰해야할지, 이를 위한 최적의 방법은 무엇인지가 포함되어야 한다.

1). 분석의 목적과 변수가 무엇이 있는지 확인. 개별 변수의 이름이나 설명을 가지는지 확인.

2). 데이터를 전체적으로 살펴보기 : 데이터에 문제가 없는지 확인. head나 tail부분을 확인, 추가적으로 다양한 탐색. (이상치, 결측치 등을 확인하는 과정)

3). 데이터의 개별 속성값을 관찰 : 각 속성값이 예측한 범위와 분포를 갖는지 확인. 만약 그렇지 않다면, 이유가 무엇인지를 확인해 본다.

4). 속성 간의 관계에 초점을 맞추어, 개별 속성 관찰에서 찾아내지 못했던 패턴을 발견한다. (상관관계, 시각화 등)

2. 이상값을 찾아내는 방법

이상치부분에서 우리가 해야하는것은 먼저 이상치가 왜 발생했는지 의미를 파악하는 것이 중요하다. 그리고 그러한 의미를 파악했으면 어떻게 대처해야 할 지(제거, 대체, 유지 등)를 판단해야한다. 이상치를 발견하는 기법은 여러가지가 있고 대표적으로 아래와 같은 방법

들이 있다.

1) 개별 데이터 관찰

데이터값을 눈으로 쭉 훑어 보면서 전체적인 추세와 특이사항을 관찰 할 수 있다. 데이터가 많다고 앞부분만 보면 안되고, 패턴이 뒤에서 나타날 수도 있으므로 뒤 or 무작위로 표본을 추출해서 관찰한다. 단, 이상값은 작은 크기의 표본에 나타나지 않을 수 있다.

2) 통계값 활용

적절한 요약 통계 지표(summary statistics)를 사용할 수 있다. 데이터의 중심을 알기 위해서는 평균(mean), 중앙값(median), 최빈값(mode)을 사용할 수 있다. 데이터의 분산도를 알기위해서는 범위(range), 분산(variance)을 사용할 수 있다. 통계 지표를 이용할 때는 데이터의 특성에 주의해야 한다. 예를 들어, 평균에는 집합 내 모든 데이터 값이 반영되기 때문에, 이상값이 있으면 값이 영향을 받지만, 중앙값에는 가운데 위치한 값 하나가 사용되기 때문에 이상값의 존재에도 대표성이 있는 결과를 얻을 수 있다. 회사 직원들의 연봉에 대해서 평균을 구하면, 대개 중간값보다 훨씬 높게 나오는데, 그것은 몇몇 고액연봉자가 평균을 끌어올렸기 때문이다.

3) 시각화 활용

일단은 시각적으로 표현이 되어있는 것을 보면, 분석에 도움이 많이 된다. 시각화를 통해 주어진 데이터의 개별 속성에 어떤 통계 지표가 적절한지 결정할 수 있다. 시각화 방법에는 확률밀도 함수, 히스토그램, 점플롯(dotplot), 워드 클라우드, 시계열 차트, 지도 등이 있다.

4) 머신러닝 기법 활용

대표적인 머신러닝 기법으로 K-means를 통해 이상치를 확인 할 수 있다.

5) 그 외의 이상치 발견 기법들

이상치 찾는 기법	예시		
Statistical-based Detection	Distribution-based, depth-based	Deviation-based Method	Sequential exception, OLAP data cube
Distance-based Detection	Index-based, nested-loop, cell-based, local-outliers		

[표1] 그 외의 다양한 이상치 발견 기법들

3. 속성 간 관계 분석하기

이 과정의 목표는 서로 의미 있는 상관 관계를 갖는 속성의 조합을 찾아내는 것이다. 여기서 부터 사실상 본격적인 탐색적 분석이 시작된다. 분석의 대상이 되는 속성의 종류에 따라, 방법도 달라져야 한다. 아래의 [표2]은 데이터의 종류를 보여준다.

Categorical Variable (Qualitative)	Nominal Data	원칙적으로 숫자로 표시할 수 없으나, 편의상 숫자화. (순위의 개념이 없음)

		예시) 남자-0, 여자-1
	Ordinal Data	원칙적으로 숫자로 표시할 수 없으나, 편의상 숫자화. (순위의 개념이 있음) 예시) 소득분위 10분위 > 9분위 > 8분위
Numeric Variable (Quantitative)	Continuous Data	데이터가 연속량으로서 셀 수 있는 형태. 예시) 키 - 166.1cm
	Discrete Data	데이터가 비연속량으로서 셀 수 있는 형태 예시) 자식 수 5명

[표2] Qualitative Data vs Quantitative Data

데이터 조합	요약통계	시각화
Categorical - Categorical	교차테이블	모자이크 플롯
Numeric - Categorical	카테고리별 통계값	박스 플롯
Numeric - Numeric	상관계수	산점도

[표3] 데이터 조합 별 통계 및 시각화 방법

1) Categorical - Categorical

교차테이블, 모자이크 플롯을 이용해 각 속성값의 쌍에 해당하는 값 개수를 표시할 수 있다.

2) Numeric - Categorical

각 카테고리별 통계값(평균, 중간값 등)을 관찰할 수 있다. 이를 박스플롯을 통해 시각적으로 표현할 수 있다.

3) Numeric - Numeric

상관계수를 통해 두 속성 간의 연관성을 나타낼 수 있다. -1은 두 속성이 반대 방향으로 변하는 음의 상관관계를 나타낸다.

0은 상관관계 없음을 나타낸다.1은 두 속성이 항상 같은 방향으로 변하는 양의 상관관계를 나타낸다. 상관계수를 갖는 두 속성의 관계도 다양한 양상을 띌 수 있는데, 스케터플롯을 이용하여 이를 시각적으로 표현할 수 있다.

또, 분석을 하다보면, 2개 이상의 속성 간의 관계를 보고싶을 때가 있다.그럴땐 위에서 나타낸 그래프를 3차원으로 표현하거나, 그래프 위에 표현된 점을 색상을 이용하거나 모양을 달리하여 더 많은 속성을 나타낼 수 있다. 혹은 각 점을 텍스트로 표현할 수도 있을 것이다.

4. 실습

실습은 Kaggle에 올라와 있는 Titanic데이터 (https://www.kaggle.com/c/titanic)를 가지고 진행한다. 먼저 데이터의 설명과 이것으로 무엇을 할 것인지 목적을 확인 한다.

Objective

승객의 정보를 보고 타이타닉에서 생존했을 지 아닐 지를 분류.

Variable Description

변수이름	설명
Survived	1 : 생존, 0 : 사망
Pclass	1 : 1등석, 2 : 2등석, 3 : 3등석
Name	승객 이름
Sex	승객의 성별
Age	승객의 나이
Sibsp	함께 탑승한 형제 또는 배우자의 수
Parch	함께 탑승한 부모 또는 자녀의 수
Ticket	티켓 번호
Fare	티켓 요금
Cabin	선실 번호
Embarked	탑승한 항구

[표3] 데이터 조합 별 통계 및 시각화 방법

목적은 Survived라는 변수를 1과 0으로 잘 분류하는게 목적이고, 이를 위해서 우리는 고객의 개인정보(이름,성별,나이 등)와 그 외 정보(티켓 번호, 요금, 선실 번호 등)들을 알고 있다.

```
df_train.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

[그림1] 데이터의 맨 위 5개의 관측치(head)

```
df_train.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

[그림2] 데이터의맨 아래 5개의 관측치(tail)

이런 간단한 방법으로 확인할 수 있는 점은 여러개가 있다.

1. Name의 경우 정해진 형식이 없이 누구는 괄호가 있고, 누구는 ""이 있는 것을 확인할 수 있다.

2. Sex의 경우 문자형태로 데이터가 주어져서 Encoding이 필요하다.

3. Ticket의 경우 숫자로만 이루어진값과 알파벳이 섞인것이 있다.

4. Cabin과 Age에서 결측치 (NaN)이 보인다.

5. Embarked의 값이 S,C,Q,가 보이는데 어떠한 항구를 의미하는지 확인해야 한다.

이런점을 기억해놓고, 시각화 및 이상치, 결측치에 대해서 자세히 탐색을 시작해야 한다. 위의 점들을 확인 한 다음에는 본격적으로 이상치와 결측치, 시각화에 대해서 분석해야 한다. 결측치는 다음주의 자료에서 자세히 다루니 패스하고 여기에서는 이상치와 시각화에 대해서 보겠다.

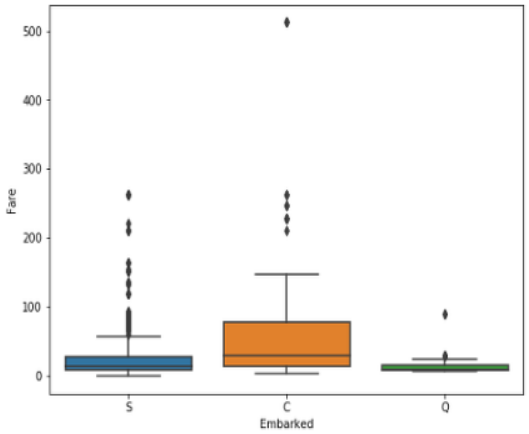
```
df_train['Fare'].describe()
```

```
count    891.000000
mean      32.204208
std       49.693429
min        0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: Fare, dtype: float64
```

[그림3] Fare(탑승 요금)의 통계값

위와 같이 통계적인 기법으로 이상치를 확인할 수 있다. 그런데 통계적인 이상치의 단점은 대부분 단변량 분석을 하기 때문에 다른 변수들간의 관계에 따라서 파악하기 어렵다는 점이다. 그래서 Fare와 밀접한 연관이 있을만한 변수인 탑승장소를 같이 시각화 하면 아래의 [그림4]와 같다.

```
data = pd.concat([df_train['Fare'], df_train['Embarked']], axis=1)
f, ax = plt.subplots(figsize=(8, 6))
fig = sns.boxplot(x='Embarked', y="Fare", data=data)
```



[그림4] Fare과 Embarked의 boxplot

위의 그림을 보면 각각의 탑승역에 따라서 이상치가 나타나는것을 확인할 수 있고, 특히나 C에서의 이상치는 다른 값들과는 달리 멀리 떨어져있는것을 확인할 수 있다. 이러한 이상치가 생긴 이유를 좀 더 자세히 보기위해 살펴보면 아래 [그림5]와 같다.

```
df_train[df_train['Fare']>500]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
258	259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3292	NaN	C
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3292	B51 B53 B55	C
737	738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3292	B101	C

[그림5] Fare이 500이상인 이상치

눈에 띄이는 특징은 나이가 비슷하고, Ticket번호가 똑같다는 점이

다. 아마 친구사이끼리 같이 티켓을 신청한것으로 생각되고, PC17755라는 티켓 자체가 VIP석을 의미하는것으로 생각된다. (Pclass가 1인것을 봐서는 1등급에서도 급이 나뉘는것 같다.)

참고자료

1. http://alphahackerhan.tistory.com/12

2. http://hellotheresy.tistory.com/26

3. https://statkclee.github.io/ml/ml-eda.html

4. https://medium.com/mighty-data-science-bootcamp/eda-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%84%A4%EB%AA%85%EC%84%9C%EC%97%90%EC%84%9C-%EC%8B%9C%EC%9E%91%ED%95%98%EA%B8%B0-230060b9fc17