**[빅데이터분산컴퓨팅 프로젝트]**



**뉴욕 택시 빅데이터를 분석을 통한**

**불쾌지수와 택시 팁의 관계**

|  |  |  |
| --- | --- | --- |
| **팀명** | 야 너둡 하둡 할 수 있어 | |
| **팀원** | **성명** | **학과/학부** |
| 박승 | 기계공학부 |
| 강승표 | 기계공학부 |
| 안민하 | 소프트웨어학부 |
| 손민하 | 경제학과 |
| **제출날짜** | 2022-12-20 | |

목 차

**제 1 장 서 론**

* 1. 주제 선정 배경
  2. 프로젝트 목적

**제 2 장 본 론**

2.1 데이터 정제

2.2 데이터 분포도

2.3 카이제곱 검정 및 회귀분석

**제 3장 결 론**

3.1 데이터 분석 결과

3.2 결론

3.3 참고문헌

**제 4장 후 기**

제 1 장 서 론

* 1. **주제 선정 배경**

미국 뉴욕시는 2010년 기준 817만 5,133명으로 미국에서 인구수가 가장 많은 도시이다. 뉴욕에 택시들은 경매를 통해 면허(Medallion)를 거래하는 형식이며 많은 인구수에 비해 면허의 공급 수량에 제한을 둬서 택시기사들의 안정적인 생활을 기대할 수 있었다.[1] 하지만 이런 면허 공급통제로 인해 2000년대부터 면허의 가격이 상승하게 되었다.[2] 특히 법적 계약서를 이해하지 못한 기사들은 면허 대상 대출을 전문으로 하는 신용협동 조합 등 금융기업들의 홍보만을 믿고 원금도 갚지 못하는 약탈적 대출을 받아왔다. 2014년 면허의 개당 가격은 약 130만달러(한화 약 14억 3천만원) 이상으로 상승하였고 이는 2014년 택시기사들의 연평균 소득이 3만 2천달러(한화 약 3천3백만원)정도임을 감안했을 때, 면허의 가격에 버블이 있음을 의미한다.[3] 2011년에는 우버와 같은 차량공유서비스가 등장하면서 면허의 가격이 2019년까지 약 90% 하락하여 택시시가들의 생계가 어려워졌다. 또 택시기사들은 임금문제 뿐만 아니라 장시간 노동으로 인한 건강문제에 시달리는 등 수준 높은 삶의 질이 보장되지 않았다. 제도적으로 택시 기사들의 처우 개선이 해결되지 않는다면 임금을 보충해줄 수 있는 또다른 수단인 팁의 개선 방법을 빅데이터 기반으로 찾기 위해 주제를 선정하게 되었다.

**1.2 프로젝트 목적**

택시 기사들은 운행료 외에 팁을 받고 있는데 팁은 서비스 제공자의 낮은 임금을 보충한다는 점에서 큰 의의가 있다. 관례적으로 요금의 10~20%의 팁을 지불하는 것으로 알려져 있으나 주관적인 부분이 개입되어 올바른 팁을 제공받고 있는지 확인하기 어렵다. 소비자의 지출은 고객의 기분이나 온도, 습도 등의 날씨에 영향을 받으며 온도, 습도와 같은 날씨는 인간의 행동과 기분에 영향을 미칠 수 있다.[4] 이에 본 프로젝트에서는 날씨에 따라 인간이 느끼는 불쾌감의 정도를 기온과 습도를 조합해 나타내는 수치인 불쾌지수와 승객의 팁 지불 비율에 대한 빅데이터 분석을 시행한다. 이를 통해 팁을 높일 수 있는 변수를 찾게 된다면 이를 고려한 맞춤 서비스 개선을 통해 택시기사들의 소득 증가, 높은 서비스 질을 기대할 수 있다.

제 2 장 본 론

**2.1 데이터 정제**

올바른 팁 비율 계산과 분석 모델을 구축하기 위해 데이터 결측치, 이상치를 제거하였다. 결제 방법의 경우 카드 결제 외에는 팁 비율의 높은 불확실성이 예상되기 때문에 Payment\_type이 Credit card인 데이터만을 사용하였다. 데이터를 아래 **Table 1** 처럼 전처리 후에 데이터는 4561485개로 확인된다. 분석을 위한 팁 비율을 계산하기 위해 Total\_amount와 tip\_amount 변수를 사용했다.

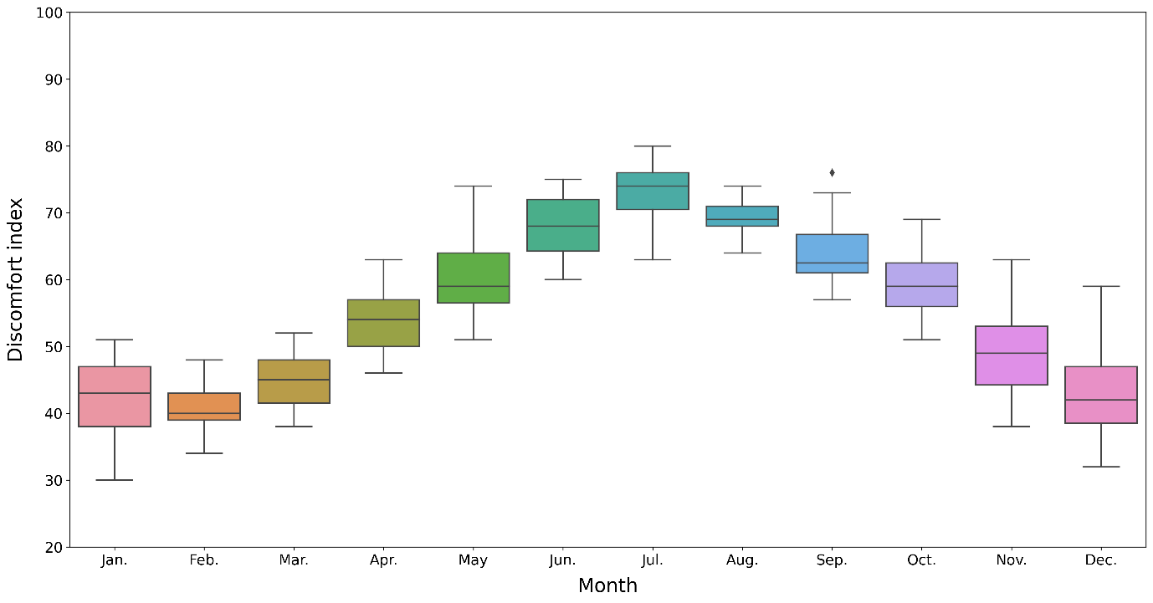
**Table 1** Taxi Data Preprocessing

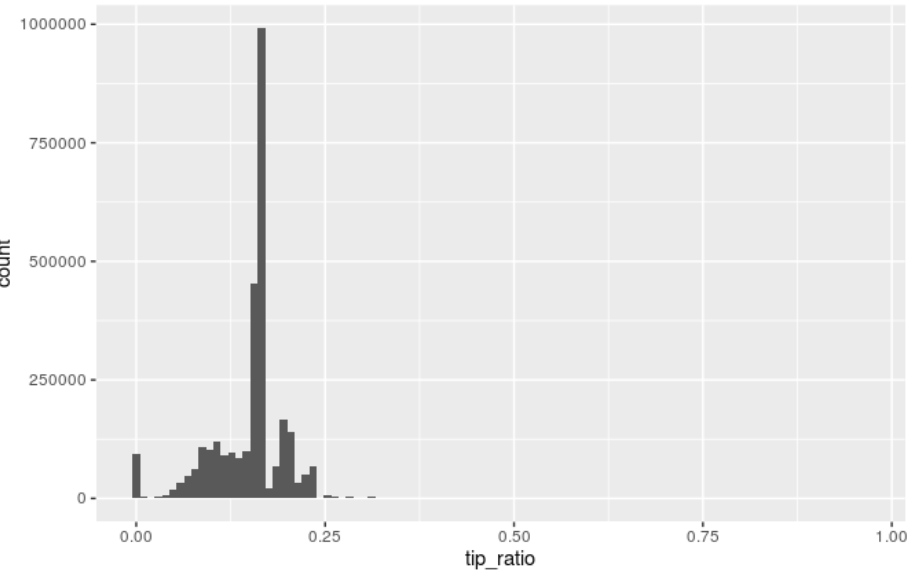
|  |  |
| --- | --- |
| Payment\_type | Payment\_type 🡪 Credit card |
| Fare\_amount | Fare\_amount > 2.5 dollars |
| Rate\_code | 0 < Rate\_code < 7 |
| Store\_and\_fwd\_flag | Store\_and\_fwd\_flag 🡪 Drop out |
| Passenger\_count | Passenger\_count > 0 |
| Trip\_time\_in\_secs | 0s <=Trip\_time\_in\_secs <= 10000s |
| Trip\_distance | 0mile <= Trip\_distance <= 200mile |

**2.2 데이터 분포도**

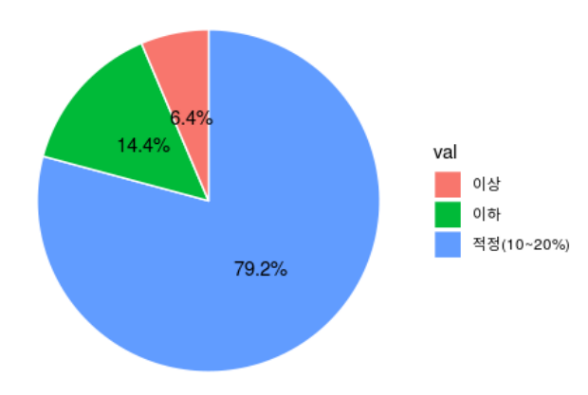
소비자의 심리 영향을 파악할 수 있도록 불쾌지수 지표를 선택하였다. 불쾌지수는 날씨에 따라 사람이 느끼는 불쾌함을 수치화해 나타낸 것으로 기온과 습도를 이용해 계산한다. 뉴욕시의 온도와 습도는 지역의 일기 예보, 기상보고서를 제공해주는 WEATHER UNDERGROUND 사이트를 참고하였다. 불쾌지수 식은 아래와 같다.

불쾌지수

**Fig. 1** 월별 불쾌지수 분포 



**Fig. 2** 팁 비율의 분포

**Fig. 1** 에서 확인할 수 있듯이 뉴욕시 온도와 습도를 바탕으로 월별 불쾌지수를 나타내었다. 일반적으로 불쾌지수가 70이상일 때 더위에 의해 불쾌감을 느끼고 50이하일 때 약간의 쌀쌀함과 상쾌감을 느낀다고 한다. 일반적으로 온도와 습도가 높은 여름철에 불쾌지수가 높다. 예상가능한 부분이지만 본 프로젝트에서는 불쾌지수와 팁의 관계를 우선시했기 때문에 불쾌지수 데이터의 분포 확인을 위함에 있다. **Fig. 3** 는 팁 비율의 분포를 나타내고 있다. 택시 운전기사에게 전체 요금의 10% ~ 20% 사이의 팁을 지불하는 것이 관례이며, 분포 상으로 봤을 때도 이러한 10%~20% 팁 비율을 대부분 이행하고 있음을 알 수 있다. **Fig. 2**를 통해 이상치 및 결측치 제거 후에도 비율이 0%인 부분이 차지한 정도를 보면 팁을 지불하지 않는 경우도 있다는 것을 예상할 수 있다.

**Fig. 3** 팁 비율 분포도

**2.3 카이제곱 검정 및 회귀분석**

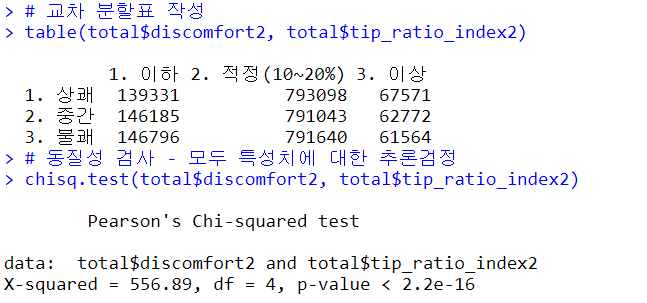
본 프로젝트에서는 이원 카이제곱 검정을 이용해 두 개 이상의 변수를 대상으로 교차 분할표를 사용한 검정 방법을 선택하였다. 검정방법으로 동질성 검정을 진행하였으며 검정 가설은 아래와 같다.

**귀무가설 : “불쾌지수에 따른 팁 지불 비율에 차이가 없다.”**

**대립가설 : “불쾌지수에 따른 팁 지불 비율에 차이가 있다.”**

또한 프로젝트의 추가적인 검증을 위해 map-reduce 함수로 불쾌지수가 팁 비율에 미치는 영향에 대한 단순선형회귀분석을 실시하였다.

제 3 장 결 론

**3.1 데이터 분석 결과**

**Fig. 4** 동질성 검증 결과

본 프로젝트에서 실행한 검증 결과를 **Fig. 4**에 나타내고 있다. 결과에 의하면 유의 확률(p-value: 2.2\*e^-16)이 0.05 이하이기 때문에 유의미한 수준에서 귀무가설을 기각할 수 있다. 따라서 “불쾌지수에 따른 팁 지불 비율에 차이가 없다.”라는 귀무가설을 기각할 수 있기 때문에 불쾌지수에 따라 승객이 지불하는 팁의 비율에 유의한 차이가 있는 것으로 해석할 수 있다.

프로젝트의 추가적인 검증을 위해 map-reduce함수로 불쾌지수가 팁 비율에 미치는 영향에 대한 단순선형회귀분석을 실시하였다. 불쾌지수를 설명변수로 갖는 단순선형회귀모형으로 결과값을 구했으며 결정계수 R2는 0.99로 1과 가까워 추정된 회귀식이 관측값들을 잘 설명하고 있다. 하지만 설명력이 매우 높게 나온 부분은 고찰할 필요성이 있다. 결정계수가 1에 가까울 수록 해당 모형이 데이터의 분포를 잘 설명한다고 할 수 있다. 그만큼 예측변수를 잘 설명하는 것으로 보이지만 샘플 및 예측 변수의 수가 R2 값을 불필요하게 증가시킬 수 있음에 주의해야 한다.

**3.2 결론**

본 프로젝트에서는 택시기사들의 면허(Medalion)가치 하락과 임금, 처우 개선을 위해 팁 소득 증대, 서비스 개선을 목적으로 하였다. 온도, 습도에 의해 고객의 소비 욕구가 달라진다는 사전조사를 통해 불쾌지수와 팁의 연관성을 분석하였다. 불쾌지수는 각국의 기상기관에 따라 모델이 다르며 본 프로젝트에서는 미국 시카고 대학의 기후학자 톰이 제안한 모델을 사용했다. 회귀 분석 모델로 카이제곱 검정을 통해 동일 집단의 두 변수를 대상으로 동질성 검증을 진행하였다. 귀무가설과 대립가설을 세웠고 유의확률 (p-value: 2.2\*e^-16)이 0.05 이하이기 때문에 불쾌지수에 따라 승객이 지불하는 팁의 비율에 유의한 차이가 있는 것을 입증할 수 있었다.

날씨와 계절에 따른 불쾌지수는 승객의 팁 지출 비율에 유의한 영향을 미친다. 따라서 이를 고려하여 택시 실내 온도와 습도를 관리하여 고객에게 높은 서비스를 제공한다면 팁으로 인한 수익을 높일 것으로 기대된다. 하지만 쾌적한 실내를 위해 공조시스템을 장시간 가동하는 것은 또다른 에너지 문제를 야기한다. 여름에 강한 햇볕으로 열기가 차량내에 전달되어 에어컨 효율을 낮춘다는 사례가 있다. 이에 단열과 자외선 차단 기능이 뛰어난 ‘열차단 필름’ 설치를 통해 공조시스템의 에너지 효율 문제를 해결하고 기여하고 승객들에게도 높은 서비스 질을 제공할 수 있을 것이다.

**3.3 참고문헌**

[1] 김정명. 뉴욕 시 전통 택시 산업의 상황과 변화, 그리고 관련 주체들의 경제적 이익.국제

노동브리프, 2020, 85-93.

[2] SNEAD, Jason. Taxicab medallion systems: Time for a change. Washington, DC: The Heritage

Foundation. Retrieved February, 2015, 26: 2018.

[3] The New York Times(2019.5.21), “Taxi Industry Leaders Got Rich. Drivers Paid the Price.”

[4] 김지호; 채은혜; 신지연. 심리적 온도가 사회적 소비에 미치는 영향: 체화된 인지를 중심으로. *사회과학 담론과 정책*, 2015, 8.2: 187-211.

제 4 장 후 기

빅데이터분산컴퓨팅 기말 프로젝트를 통해 분산 컴퓨팅(Distributed computing)의 원리에 대해 깊이 배울 수 있었다. Map reduce 함수를 직접 다뤄보고 회귀 모형에 대해 학습하면서 대용량 데이터를 가공하고 유의미한 결과를 도출해내는데 흥미를 느꼈다. 빅데이터를 기반으로 주제를 선정할 때, 모두가 평범한 주제가 아닌 실현 가능하면서 재밌는 주제를 생각해냈고 팁과 불쾌지수의 연관성을 찾아낼 수 있었다. 팀으로 활동했기 때문에 코딩, 통계 모델, 데이터 전처리에 어려움이 생겨도 빠르게 해결할 수 있었다.

하둡을 다루면서 코드나 시스템이 익숙하지 못해 다른 프로그램을 응용해보기도 했다. 확연하게 대용량 데이터를 분석하는데 시간적으로 차이가 나고 하둡의 유용성을 느낄 수 있었다. 이 과정에서 하둡의 내장된 함수, 코드를 직접 짜보며 실력을 쌓았고 결과의 의미를 분석할 수 있었다. 결론적으로 빅데이터 분석 역량을 기르고 팀원들 모두가 발전할 수 있는 프로젝트였다.

|  |  |  |
| --- | --- | --- |
| **일시** | **내용** | **참여자** |
| 11/30~12/7  (17:00~22:00)  ZOOM | 프로젝트 주제 조사 | 박승, 강승표, 안민하, 손민하 |
| 서론 자료 조사 | 박승, 강승표, 안민하, 손민하 |
| 주제 배경 선정 | 박승, 손민하 |
| 주제 목적 선정 | 박승, 손민하 |
| 12/8~12/14  (17:00~22:00)  ZOOM | 결측값, 이상치 확인 | 박승, 강승표, 안민하, 손민하 |
| 데이터 전처리 | 강승표, 안민하 |
| 분석에 적합한 변수 확인 | 박승, 강승표, 안민하, 손민하 |
| 전처리 후 데이터 분포도 시각화 | 강승표, 안민하 |
| 분석에 사용될 이론 조사 | 박승, 손민하 |
| 12/15~12/20  (17:00~22:00)  학교 | 분석 모델 구축 | 강승표, 안민하 |
| 결과 해석 | 박승, 강승표, 안민하, 손민하 |
| 결론 도출 | 박승, 강승표, 안민하, 손민하 |
| 결과보고서 작성 | 박승, 강승표, 안민하, 손민하 |