Data scientist
개인적인 경험 및 소회

안녕하세요!
반갑습니다!

이정민

Data scientist 근로소득자 8년차

Computer science 공부 18년차

2022년 기준

Data scientist 로써
진행했던 업무, 개인적 경험과
소회를
정리해보고자 합니다

1

어떤 일들을 해왔나요?

근로소득자 Data scientist

- 전사 경영 지표 플랫폼팀에서 지표 수집, 정제, 분석 결과 제공
- 모바일 앱 사용자 로그 데이터를 이용한 사용자 쇼핑 행동 분석
- 상품 추천 모델 개발
- 기초 통계, 머신 러닝, 딥러닝 기술을 이용하여 데이터 분석, 피쳐 발굴, 모델 생성

인간 Data scientist

- 매매 할 부동산을 선택하기 위해 장소 비교를 위한 피쳐 정의
  - 직주근접, 강남 접근성, 주변 소/대형 상권, 공공시설, 편의시설 등

- 23주차에 태어난 쌍둥이 조카의 daily 몸무게 및 수유량 시계열 분석
  - 인큐베이터 체류 기간 : 약 4개월
  - 목적 : 건강하게 잘 지내고 있는지 확인하기 위해
  - 가설 : 몸무게, 수유량이 꾸준히 증가하고 있다면 건강할 것이다
  - 근데 왜? : 너무 일찍 태어난 게 걱정되서, 법적보호자가 아니라 인큐베이터 방문을 못해서, 하다보니 재밌어서

쌍둥이 조카 몸무게/수유량 시계열 분석

데이터 수집

주양육자에게 (비)주기적으로
자료를 달라고 요청

메신저를 통해 (비)주기적으로
자료를 전달 받음

데이터 정제

비정형 데이터에서
필요한 부분만 수동으로 발췌

데이터 적재

테이블 스키마 정의

저장소로 스프레드 시트 선택

daily 기준으로 테이블에 값 입력

추가 지표 발굴

조카 A와 조카 B의 몸무게 차이

지표 시각화

Line chart

X 축 : 날짜

Y 축 : 몸무게, 수유량

차트 해석 및 결론 도출

꾸준히 증가하는 추세이다

몸무게 차이가 크게 벌어지지 않는
경향성을 보인다

건강하게 잘 자라고 있는 것으로 보인다

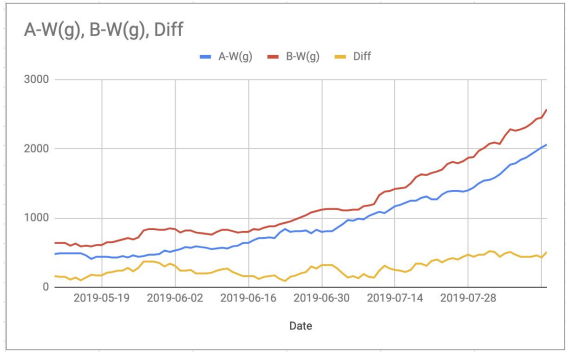
쌍둥이 조카 몸무게/수유량 시계열 분석

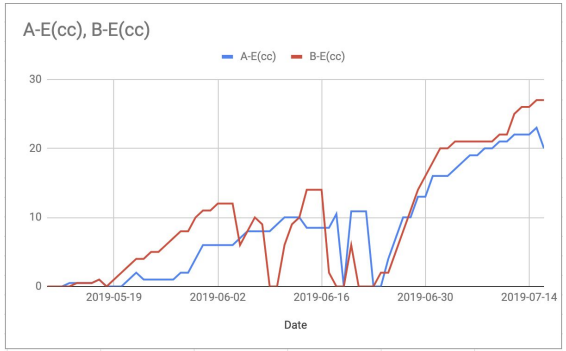
5.23.
—A. 450g. 1cc.
—B. 690g. 4cc.

5.24.
—A. 430g. 1cc.
—B. 710g. 5cc.

5.25. 26+0/생후16
—A. 460g. 1cc.
—B. 690g. 5cc.

	A	B	C	D	E	F	G	H	I
1	Date	Week	Day of week	Born	Date	A-W(g)	A-E(cc)	B-W(g)	B-E(cc)
15	2019-05-23	25	5	14	2019-05-23	450	1	690	4
16	2019-05-24	25	6	15	2019-05-24	430	1	710	5
17	2019-05-25	25	0	16	2019-05-25	460	1	690	5
18	2019-05-26	26	1	17	2019-05-26	440	1	720	6





데이터 분석은 멀리 있지 않습니다

대단히 거창하거나 어려운 것이 아닙니다

Data scientist 로써 근로소득을 발생시키는 것은 조금 다를 수 있습니다

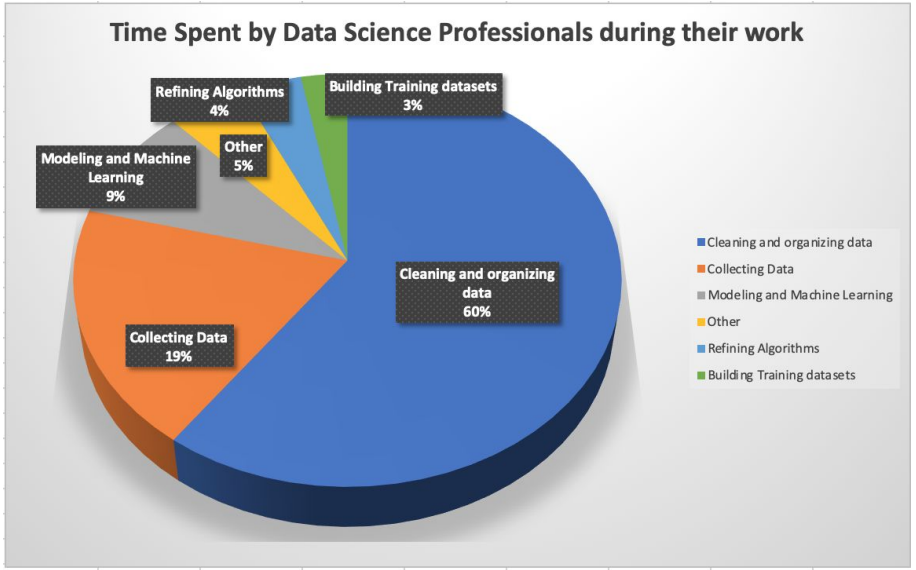
2

Data scientist 에 대한 환상

Data scientist 로써
데이터 분석을 통해
빛나는 인사이트를
찾을 수 있을거야!

데이터 과학자들이
가장 많은 시간을
보내는 업무

데이터 클렌징
(전체 업무 시간의 60%)



이미지출처

3

그럼 Data scientist 는 무엇을 하나요?

Data scientist

문제 정의
가설 수립
실험 진행
결론 도출
근데 이제 data 를 곁들인

조직/데이터 규모에 따라 다양한 업무 영역

데이터 수집

다양한 곳(web, db, file 등) 에 흩어져 있는 데이터를 한 곳으로 모으기

데이터 정제 및 적재

비정형 데이터의 정형화

버전마다 다른 로그 포맷

여러 테이블에서 필요한 컬럼만 뽑아 하나의 테이블로 모으는 것

나이 : 1, 2, 3, 4, 5 vs age : 10대, 20대, 30대, 40대 이상

기간이 다른 두 테이블을 하나로 병합

조직/데이터 규모에 따라 다양한 업무 영역

분석 지표 발굴

지표 정의, 집계 기간 정의, 선행지표, 후행지표

분자/분모 형태로 정의되는 지표의 경우 분모가 0이 되지 않도록 주의

지표가 커질 때 의미, 작아질 때 의미를 미리 해석해보자

지표의 이름을 봤을 때 사람이 기대하는 결과가 나오도록 이름을 잘 짓고, 정의를 명확하게 해보자

데이터 분석

EDA (Exploratory Data Analysis), feature analysis

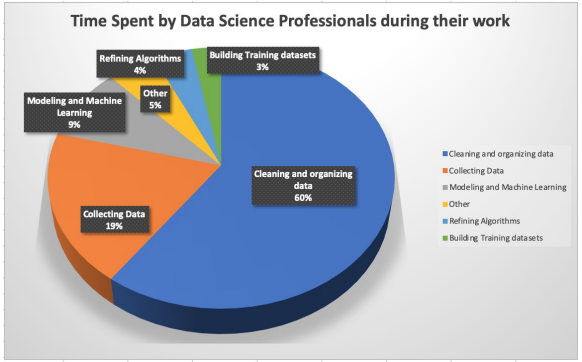
data mining, design/apply algorithm

조직/데이터 규모에 따라 다양한 업무 영역

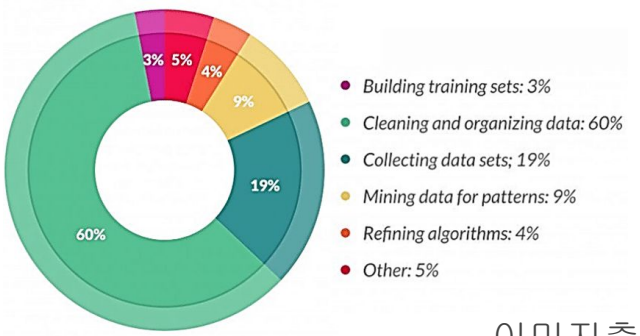
지표 시각화

시각화 하려는 종류에 따라 적절한 차트 타입 선택하기

bar chart vs line chart - 데이터 포인트 사이에 연결이 필요한가 아닌가



VS



이미지 출처

조직/데이터 규모에 따라 다양

모델 생성

모델 종류 - 예측 모델, 추천 모델, classifier, ranking 모델 등

트레이닝/테스트/validation 데이터 수집 및 정제

카테고리컬 데이터와 continuous data 구별

값이 비어있는 피쳐에 대한 디폴트 값 정의

모델 트레이닝 및 성능 측정 (필요시, 성능 측정 지표 설계)

모델 배포

모델 성능 트래킹을 위한 메트릭 정의

배포 후 모델 성능 트래킹

비즈니스 도메인에 따라 다양한 분석 주제

쇼핑
매출 분석
고객 분석
판매자 분석

게임
게임 행동 분석
사용자 결제 분석
물리 엔진 분석

경영 컨설팅
사업 분야 매출 분석
신규 사업 발굴

검색엔진
검색어 분석
사용자 분석
문서 분석 (text, image, video)

의료
질병 분석
병원 이용 고객 분석

금융
대출 분석
마케팅 타켓 분석

비즈니스 도메인에 따라 다를 수는 있지만 …

주로 소통하게 될 사람은 의사 결정권자
의사 결정에 확신을 얻고 싶어한다
미래를 대응할 수 있는 실현 가능한 구체적 방안을 원한다
데이터의 단순 현재 상태를 보고 받고 싶어 하는 의사결정권자는 없다

분석 결과를 전달한 사람 입장에서 들을 수 있는 황당하면서도 최악인 말
그래서 어쩌라고?
어떻게든 결론을 내보려 노력해야 한다

4

매출 분석 결과 좀 봅시다

가상 시나리오

1월 매출은 x 억원, 2월 매출은 y 억원, 3월 매출은 z 억원입니다

1월 매출은 x 억원, 2월 매출은 y 억원, 3월 매출은 z 억원으로 증가 추세이며,
YoY 대비 추세 역시 약 10% 정도로 증가 추세에 있습니다
이러한 추세가 지속된다면 4월 예상 매출은 w 억원입니다

추세 분석
구체적 수치를 덧붙임
예상 매출 제시

예상 매출을 증대를 위해, 과거에 진행했던 쿠폰 이벤트를 기반으로 분석한 결과입니다
과거 쿠폰 이벤트에 투입한 비용은 일주일 기준 x 억원이며,
당시 쿠폰을 직접 사용한 매출은 y 억원입니다
만약, 이번에도 비슷한 비용을 투입한다면 예상 매출은 z 억원입니다

매출 증대라는 매력적인 목적(가설 - 쿠폰 이벤트를 하면 매출이 늘어날 것이다) 제시
과거 데이터를 기준으로 분석
  기존에 효과를 입증한 이벤트를 기반으로 예측
예상 매출 제시

1년 중 매출이 집중되는 달 중의 하나인 5월 매출 증대를 위해,
4월 한달 동안 신규 사용자를 늘리는 것을 제안합니다
신규 사용자 증가 추세가 현재 정체 상태로 보이기 때문에,
4월 한달 동안 신규 사용자를 늘리는 것이 5월 매출 증대를 위한 방안이 될 수 있습니다
신규 사용자의 구매 패턴 분석 결과를 보면, 가입 후 1~2달 사이에 신규 주문을
발생시킨 경우, 그 이후에도 꾸준히 주문하는 경향이 있는 것으로 보입니다.
따라서 신규 사용자의 4월 가입 및 첫 주문을 완료하도록 하는 것이 중요합니다

매출 증대를 위한 새로운 방안 제시 (신규 사용자 모집)
새로운 방안이 필요한 타당한 근거 제시 (5월이 중요한 이유, 사용자 증가 추세가 정체)
새로운 방안이 정말 효율적인지에 대한 근거 제시 (신규 사용자 구매 패턴 분석 결과)
새로운 방안이 실제 매출 증대로 이어지기 위한 구체적인 목표 제시 (가입 및 첫 주문 완료 유도)

요구 사항 구체화는 어떻게 할 수 있을까?

끊임 없이 질문을 던져야 한다
전체 매출의 추이를 보고 싶은 것인가?
상품 카테고리 별 매출 추이가 필요한가?
사용자 그룹 별 매출 추이가 필요한가?
YoY, QoQ 대비 매출 추이가 필요한가?
반복적이고 정형화된 요구 사항은 자동화/시각화로 대체할 수 있다

의사결정권자의 입장이 되어서 생각해보자

5

Data scientist 로써 일을 잘하려면 무엇을 알아야 하나요?

기본 지식

Computer science
수학, 통계
자동화 도구
비즈니스 목표
비즈니스 도메인의 특징
커뮤니케이션 능력

직무적인 스킬

SQL은 기본
직접적으로 SQL statement를 쓰지 않더라도 sql function에 대한 이해는 필수
알지만 쓰지 않는 것과, 몰라서 못 쓰는 것은 같지 않다

기본 통계는 의외로 중요하다
count, distinct count, avg, mean, std, max, min
분석하려는 데이터의 대략적인 큰 그림을 짚고 넘어가야 한다

직무적인 스킬

분석하려는 데이터의 분포를 아는 것이 필요하다
각 값의 범위는 어떻게 되는지
outlier가 많은 데이터인가

join 이전/이후 count 비교는 중요하다
join 이후 불필요한 데이터 중복이 발생하지 않았는지 습관적으로 확인하라
join 하려는 key 분포가 skewed 되어 있는지 확인하자

직무적인 스킬

개발자에게 프로그래밍 언어는 요리사가 쓸 수 있는 조리도구와 비슷하다

많은 언어를 다룰 수 있는 사람은, 그렇지 않은 사람과 다르다

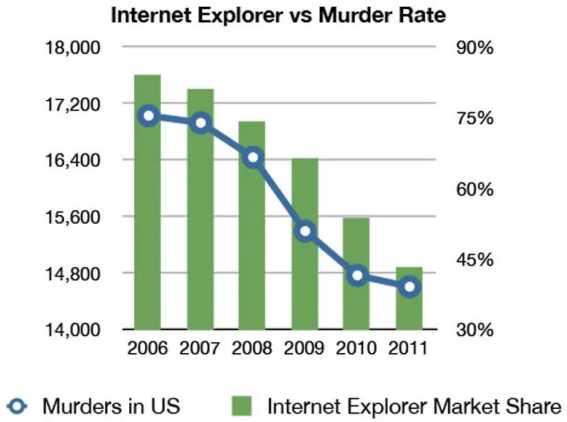
하나의 언어만을 다루더라도 깊숙한 기능까지 다룰 수 있는 사람은,
그렇지 않은 사람과 다르다

나만의 주 무기 하나를 장착하는 것은 중요하다

주 무기 외에 보조 무기까지 있다면 금상첨화 일 수 밖에 없다

Correlation vs causation

상관관계가 있다고 해서 그것이 인과관계의 성립을 의미하지 않는다



Using Internet Explorer leads to murder

차트 출처
The 10 Most Bizarre Correlations

추천 문서
담뱃값 오르면, 흡연율 낮아질까?

6

신입 data scientist 에게 전하고 싶은 말

Morality

황우석 사태를 알고 있는가 ?

데이터 조작은 할 수 있다 하더라도, 절대 해서는 안 될 행동이다

원하는 결론을 도출하기 위해 입맛에 맞게 지표를 정의하고,
필요한 데이터만 선별하는 행동은 하면 안된다

개인 정보 보호도 항상 신경써야 한다

나쁜 질문은 없다

업무 진행에 꼭 필요한 질문이라고 생각된다면 적극적으로 질문해보자

단어 하나 하나의 의미, 축약어의 의미 등등
궁금한 것이 매우 많을 수 있으며, 그것은 아주 자연스럽다

신입이라는 방패를 적극적으로 활용하자

Easy come, easy go

나쁜 질문은 없지만, 더 좋은 질문은 있다
질문하기 전에 스스로 답을 구하려는 노력은 중요하다
동료에게 물어봐서 쉽게 얻은 지식은 쉽게 사라진다
    이거 어떻게 하는지 아시나요?
    제가 이렇게 저렇게 시도해봤는데 안되네요
기억보다는 기록을 믿자

꾸준히 공부해야 한다

아무것도 공부하지 않으면, 자연스럽게 정체되고 도태된다

새로운 기술/지식에 대한 트래킹은 꾸준히 하되,
그것을 실무에 반영하는 것은 시기와 ROI를 잘 따져봐야 한다

우리가 만들어야 할 것은, 먹을 사람의 취향에 맞는 맛있는 음식이다

그것을 위해 매번 제철 식재료와 값비싼 요리도구가 필요한 것은 아니다

기초 통계 지표만으로 데이터 분석 업무를 진행한다고 해서 자괴감을
느끼거나 실망할 필요도 없다

회사라는 조직에서 일을 한다는 것

개인의 커리어는 개인이 챙겨야 한다
이력서를 주기적으로 업데이트 해보자
이력서에 쓸 말이 없다면 스스로의 상태를 점검해 볼 필요가 있다
본인이 진행한 업무를 어떻게든 이력서의 한 줄로 표현해보려는 노력은 필요하다

회사라는 조직에서 일을 한다는 것

개인이 열심히 업무를 진행하는 것, 업무의 성과가 좋은 것
개인의 업무 성과가 회사에서 성과로 인정 받는 것
해당 성과가 개인에게 만족스러운 보상으로 돌아오는 것

이 모든 단계는 서로 독립적인 경우가 많습니다
스스로가 열심히 했지만 충분한 보상을 못 받았다고 느낄 수 있습니다
그러나, 그것은 결코 여러분이 잘못해서가 아닙니다
보상을 잘 받을 수 있는 업무는 무엇인 지 한번쯤 고민해 보는 것을 추천합니다
때로는 내 가슴이 뛰는 일을 하는 할 때 더 짜릿하기도 할 것입니다

근로소득자로써, 회사라는 조직에서 무엇을 목적으로 일할 것인지에 대해
스스로 고민해보시고 본인만의 목적을 찾기를 바랍니다

7

맺음말

지금까지 보신 내용은 다분히 저의 개인적인 경험에 의한 의견입니다
다양한 사람의 의견을 참고하시고,
본인만의 시각으로 받아들여 주시길 바랍니다

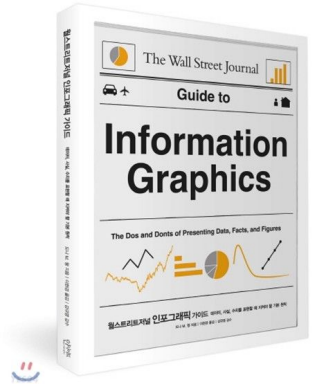
새로운 시작은 항상 설레고 기대되기 마련입니다
그러나 여러분의 앞길에는 꽃길과 가시밭길이 번갈아 나타날 것 입니다
꽃길을 당연하다 받아들이지 마시고, 가시밭길에 쉬이 절망하지 않으시길 바랍니다

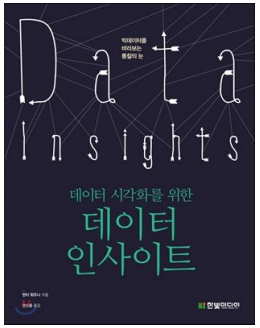
감사합니다!

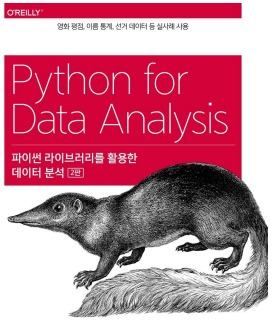
질문 있으세요?

You can find me at LinkedIn

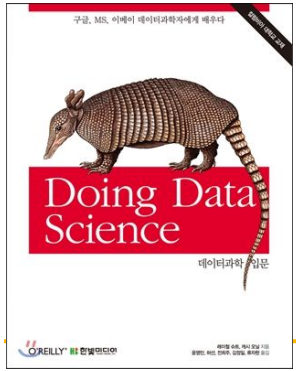
추천 도서


The Wall Street Journal
Guide to
Information Graphics
The Dos and Donts of Presenting Data, Facts, and Figures


Data Insights
데이터 시각화를 위한
데이터 인사이트


O'REILLY
Python for Data Analysis
파이썬 라이브러리를 활용한 데이터 분석 2판


O'REILLY
파이썬 데이터 사이언스 핸드북


구글, MS, 이베이 데이터과학자에게 배우다
Doing Data Science
데이터과학 입문
O'REILLY