1. 研究背景

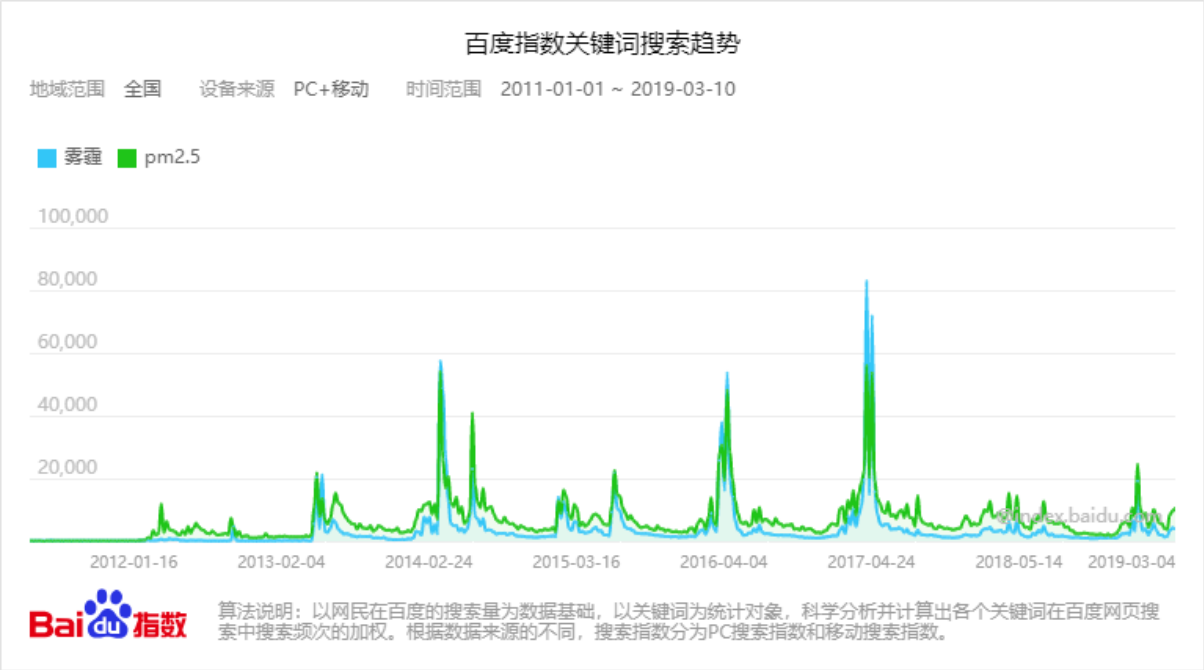

图 1 雾霾和 PM2.5 的百度搜索指数

2011 年左右开始，美国大使馆持续播报 PM2.5 的行为终于引发了社会关注。2013 年，“雾霾”成为年度关键词，自此之后，社会包括政府都开始对空气质量给予高度的关注。2012 年，中华人民共和国环境保护部发布了《环境空气质量指数（AQI）技术规定（试行）》并于 2016 年 1 月 1 日期正式实施。

城市交通是与市民生活息息相关的一部分，随着我国汽车保有量的增多，与大城市较为集中的出行时间，交通拥堵成为了市民经常会在社交媒体上讨论的事情。

2. 研究意义

为了降低我国空气污染，提升空气质量，改善人们的生活水平，很多专家已经从各个行业领域提出了可能的解决方案，如增加城市绿化水平，对工业污染进行处理限制其直接排放等。在本研究中，我们希望通过探究空气质量与交通各个方面的相关性，探究空气质量是否会与交通的某个方面有着较为显著的相关性，并根据其之间的关系提出一种新的减轻城市空气污染或交通问题的可能性，同时尝试能否从中提取出一定的商业价值。

3. 团队成员及任务分工

全国车辆事故分析：陈力群，王洁怡，许昕，甄赏

杭州交通拥堵分析：雷鸣，杨弈，张灿，郑希民

4. 前期研究

项目最初，我们以地区划分为标准进行两部分的分析：

首先以北京，上海，广州，深圳四个城市进行交通拥堵与空气质量的相关分析，其中交通拥堵分为陆路交通拥堵及航班延误。得到北上广深四城市的 AQI，PM2.5 和 PM10 的指数趋势以及各城市的年度污染等级和空气污染物比例分布，对相关趋势及分布进行比较分析，发现北京，上海两城市空气污染程度较为严重，并且 PM2.5 及 PM10 为主要污染物。通过 Python 网络爬虫得到的四城市日度交通拥堵的百度搜索指数，并与四城市的日度 AQI 实际指数进行趋势对比分析。综合四市相较，当空气质量指数越高即污染越严重时，交通拥堵搜索指数相应会上升，且搜索指数于三四季度有明显峰值出现。整合在飞常准大数据平台上获得的四城市月度航班准点率，发现四城市航班准点率在第二三季度较低，北京上海航班准点率较于广州深圳偏低，北京航班准点率与 AQI 指数大致呈反向变化，其余三市航班准点率与 AQI 指数的反向变化呈偶发性。

其次以华中，华东，华南，华北四个地区进行交通事故与空气质量的相关分析，其中交通事故由公布的已知相关报告汇总得到月度总数。通过对四地区的空气质量进行初步分析，得到华东华北地区空气污染情况比华中华南地区显著，且重度污染发生率较高，尤其发生在第一第四季度。通过对各地区的交通事故发生数目与 AQI 指数进行对比分析，得到华东华南地区交通事故发生数目较华中华北地区略高，且与 AQI 指数趋势大致相同。

5. 空气质量和全国交通事故的影响分析

5.1. 数据收集

在本部分，所有的数据可以大致的分为三个部分：空气质量相关数据的实际数值和百度搜索指数，“车险”、“车险理赔”和“车载空气净化器”这三个关键词的百度搜索指数。其中，空气质量数据包含了空气质量指数（AQI）、PM2.5、PM10、二氧化氮（NO2）、二氧化硫（SO2）、一氧化碳（CO）和臭氧（O3）。“车险”、“车险理赔”和“车载空气净化器”这三个关键词的百度搜索指数是运用 python 网络爬虫爬取的 2015 年至 2018 年共四年的数据。

5.2. 数据说明

有 2015 年 1 月 1 日到 2018 年 12 月 31 日共 1461 条日度数据。包括详细日期和按城市人口加权平均后的全国空气质量指数日度数据以及百度搜索指数日度数据。

表 1 数据说明表

变量类别	类型	具体变量	时间	取值范围
空气质量	数值型	AQI、PM2_5、 PM10、 SO2、 CO、 NO2、 O3_8h	2015-2018 年	大于 0
百度指数	连续型	“车险”、”车险理赔”和“车载空气净化器”	2015-2018 年	大于 0

5.3. 数据分析

5.3.1. 相关性分析

通过Pearson相关性检验，对“车险”、“车险理赔”和“车载空气净化器”这三个关键词的百度搜索指数与全国空气质量指数（按城市人口加权平均）进行了相关性分析。

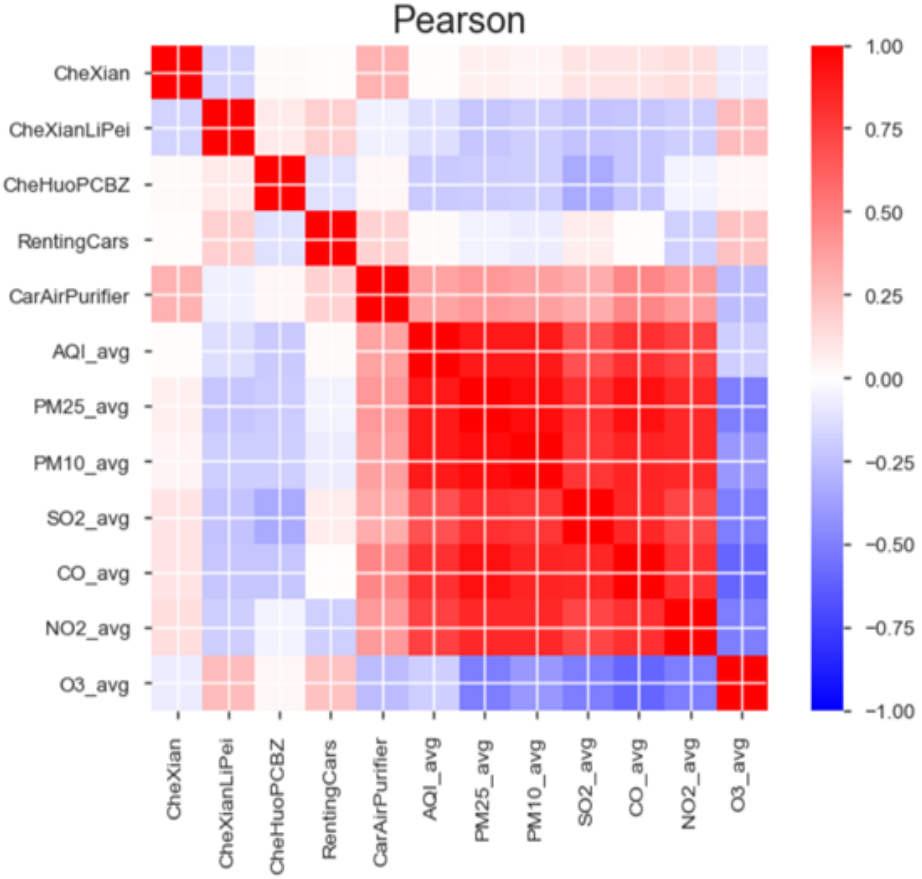

图 2 Pearson 相关性表

从相关性图中，可以得到三个初步结论：

第一，“车险理赔”与 PM2.5 平均浓度值呈现相对显著的负相关关系。

第二，“车险”与 NO2 平均浓度值呈现相对显著的正相关关系。

第三，“车载空气净化器”与 CO 平均浓度值呈现相对显著的正相关关系。

为了验证上述结论，我们采取了增加工具变量的方法，进行三次回归方程的拟合，并利用 Kappa 检验维度间的共线性，接下来是具体对每一个结论的检验过程。

5.3.2. “车险理赔”与 PM2.5 平均浓度的关系

选取“车险理赔”关键词和空气质量数据，首先用 kappa 检验自变量之间的共线性值为 261.8934,根据共线性判断标准，判断为中度共线。

根据 AIC 向前逐步分析，根据解释性与共线性检验，选出 PM10 及 PM2.5 平均浓度值与各空气质量百度指数之间的线性关系，并且得到其 kappa 检验值为 37.00395，共线性降低为轻度共线。

分别选出对 PM10 及 PM2.5 百度搜索指数进行预测的最优自变量项（空气质量百度搜索指数项），得出 PM2.5 平均浓度的预测值。

从以上模型看出，变量 PM10 统计不显著，所以剔除该变量，得到新的 PM2.5 预测值。

表 2 PM2.5 回归模型系数显著性表

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	11.7426819	<2e-16***
PM2.5 百度指数	0.4025074	<2e-16***
PM2.5*PM10 百度指数	-0.0033785	<2e-16***
Adjusted R^2 = 54.75%		

PM2.5 预测值作为工具变量加入关键词分析模型，通过共线性检验和 AIC 向前选择分析，得到 kappa 值为 4.054726，判断为轻度共线。得到以下结果：

表 3 车险理赔回归模型系数显著性表

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	276.5561	<2e-16***
pre_PM2.5	-0.6052	6.60e-8***
PM10 百度指数	2.3671	<2e-16***
口罩百度指数	-1.2838	3.21e-16***
Adjusted R^2 = 12.71%		

5.3.3. “车险”与 N02 平均浓度值的关系

选取“车险”关键词和空气质量数据，首先用 kappa 检验自变量之间的共线性值为 256.1788，根据共线性判断标准，判断为中度共线。

根据 AIC 向前逐步分析，根据解释性与共线性检验，选出 AQI、O3 及 NO2 平均浓度值与各空气质量百度指数之间的线性关系，并且得到其 kappa 检验值为 9.45384，共线性降低为轻度共线。

分别选出对 AQI、O3 及 NO2 百度搜索指数进行预测的最优自变量项（空气质量百度搜索指数项），得出 NO2 平均浓度的预测值。

表 4 NO2 回归模型系数显著性表-1

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	2.840e+01	1.03e-06***
NO2 百度指数	2.404e-01	0.27003
AQI 百度指数	1.690e-01	0.00446***
O3 百度指数	-5.645e-01	0.00169***
NO2*AQI 百度指数	-6.755e-04	0.73093
NO2*O3 百度指数	5.618e-03	0.29473
AQI*O3 百度指数	1.517e-03	0.39431
NO2*AQI*O3 百度指数	-6.539e-05	0.14913
Adjusted R^2 = 38.6%		

从以上模型看除，单个变量 AQI、O3 统计并不显著，所以剔除该变量。得到新的 NO2 预测值。

表 5 NO2 回归模型系数显著性表-2

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	1.6776501	<2e-16***
AQI 百度指数	0.2587123	<2e-16***
O3 百度指数	-0.0018156	0.969
AQI*O3 百度指数	0.0036220	<2e-16***
Adjusted R^2 = 37.53%		

NO2 预测值作为工具变量加入关键词分析模型，通过共线性检验和 AIC 向前选择分析，得到 kappa 值为 4.054726，判断为轻度共线。得到以下结果。

表 6 车险回归模型系数显著性表

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	2420.2157	<2e-16***
pre_NO2	53.8448	1.5e-12***
空气净化器百度指数	26.6873	<2e-16***
pre_NO2:SO2 百度指数	-2.0908	<2e-16***
Adjusted R^2 = 23.72%		

5.3.4. “车载空气净化器”与 C0 平均浓度值的关系

选取“车载空气净化器”关键词和空气质量数据，首先用 kappa 检验自变量之间的共线性值为 263.762,根据共线性判断标准，判断为中度共线。

根据 AIC 向前逐步分析，根据解释性与共线性检验，选出 PM2.5、O3 平均浓度值与各空气质量百度指数之间的线性关系，并且得到其 kappa 检验值为 8.715446，共线性降低为轻度共线。

分别选出对 PM2.5、O3 及 CO 进行预测的最优自变量项（空气质量百度搜索指数项），得出 CO 平均浓度的预测值。

表 7 CO 回归模型系数显著性表-1

	Estimate Coefficients	**Pr(>|t|)**
(Intercept)	1.051e+00	7.99e-14***
CO 百度指数	-3.659e-03	0.0859
PM2.5 百度指数	24.635e-03	3.16e-05***
O3 百度指数	-2.152e-02	1.85e-06***
CO*PM2.5 百度指数	1.343e-05	0.3900
CO*O3 百度指数	2.445e-04	1.59e-05***
PM2.5*O3 百度指数	-1.227e-05	0.7177
CO*PM2.5*O3 百度指数	-6.854e-07	0.0673
Adjusted R^2 = 59.39%		

从以上模型看除，单个变量 PM2.5、O3 统计并不显著，所以剔除该变量。得到新的 CO 预测值。

表 8 CO 回归模型系数显著性表-2

	Estimate Coefficients	Pr(>|t|)
(Intercept)	5.833e-01	<2e-16***
PM2.5 百度指数	6.363e-03	<2e-16***
O3 百度指数	1.716e-03	0.101
O3*PM2.5 百度指数	-8.132e-05	<2e-16***
Adjusted R^2 = 57.27%		

CO 预测值作为工具变量加入关键词分析模型，通过共线性检验和 AIC 向前选择分析，得到 kappa 值为 2.84587，判断为轻度共线。得到以下结果。

表 9 车载空气净化器回归模型系数显著性表

	Estimate Coefficients	Pr(>|t|)
(Intercept)	1.048e+03	<2e-16***
pre_CO	-4.971e+02	1.67e-10
空气净化器百度指数	-6.366e+00	1.57e-08***
CO 百度指数	-1.159e+01	<2e-16***
pre_CO*空气净化器百度指数	7.373e+00	<2e-16***
pre_CO*CO 百度指数	6.710e+00	1.04e-09
空气净化器*CO 百度指数	1.164e-01	6.96e-16
pre_CO 空气净化器*CO 百度指数	-8.819-02	<2e-16***
Adjusted R^2 = 60.57%		

5. 4. 分析总结

民众对 PM2.5 的重视度相对较高，并且多次搜索关键词“口罩”，所以当 PM2.5 浓度增大时，防护意愿上升，出行意愿降低，因此“车险理赔”搜索指数下降。

现在大众对 PM2.5 的重视度相对较高，大家有各种渠道能够及时获得实时的 PM2.5 的数值。当人们发现 PM2.5 浓度有上升趋势时，往往会避免室外活动而重新规划自己的出行日期。随着出行意愿减少，道路上车辆数量减少，发生车祸的概率降低，因此发生车祸后对“车险理赔”的关注度也降低，“车险理赔”搜索指数下降。

当 NO2 浓度上升时，大气可见度降低，民众直观感受到空气质量变差，由于大气能见度降低，导致交通事故易发，所以“车险”搜索指数上升。

当 CO 浓度上升时，汽车尾气上升，民众直观感受到空气质量变差，由于汽车尾气上升，大气能见度降低，导致交通事故易发，所以“车载空气净化器”搜索指数上升。